Korkeakoulujen kielikeskus
Jyväskylän yliopisto
40100 JYVÄSKYLÄ
Puh. (tel.) 941-291 211

Language Centre for
Finnish Universities
University of Jyväskylä
SF - 40100 JYVÄSKYLÄ
Finland

FINLANCE Vol. III 1984

FINLANCE

The Finnish Journal of Language Learning and Language Teaching

VOLUME III 1984

EDITED BY LIISA KORPIMIES

Language Centre for Finnish Universities
University of Jyväskylä · Finland

F I N L A N C E

The Finnish Journal of Language Learning and Language Teaching

Vol. III 1984

Edited by

Liisa Korpimies

Language Centre for Finnish Universities

University of Jyväskylä · Finland

ISSN 0359-0933

PREFACE

The main theme of the present volume is language testing. Language testing is one of the key issues in the development of language centre teaching in Finland. Most of the articles deal with different aspects of language testing. Language testing is not, of course, an independent phenomenon. On the contrary, it is closely related to language teaching in general, to materials and methods and to the theory that supports them. The remaining articles in this volume deal with theoretical and practical aspects of language teaching.

Jyväskylä, December 1984

Liisa Korpimies

CONTENTS

Peter S. Green
University of York

Karlheinz Hecht
University of Munich

RELIABILITY IN THE ASSESSMENT OF WRITTEN COMMUNICATIVE SKILLS

In line with current trends in foreign-language teaching, the initial stages of learning English in German schools (grades 5 - 10) are concerned with "the training of basic communicative skills and abilities" (official syllabus for the *Gymnasium* in Bavaria). The problem of assessing those skills in tasks that are valid as communication and at the same time reliable as tests are well-known. The question may therefore be asked, "Are pupils being fairly assessed in the normal school tests, which are of such importance to them?"

A current research project conducted jointly by the "Lehrstuhl für die Didaktik der englischen Sprache und Literatur" of the University of Munich and the Language Teaching Centre of the University of York is looking at the performance of German pupils in written communicative skills in English and of their teachers in assessing them. This article considers particularly some aspects of the performance of the teachers as markers.

Outline of procedure

In order to test their ability to communicate in writing in English, 60 German pupils, 20 in each of the three types of Bavarian secondary school - *Gymnasium*, *Realschule* and *Hauptschule*[1], were asked to write a letter in response to a letter of elicitation. Each letter was marked and graded by three German teachers of English from the appropriate type of

[1] In the selective school system in Bavaria, the more academic pupils (those expected to remain in school to the age of 19/20) attend the *Gymnasium*, the less academic the *Realschule* (some transferring to the *Gymnasium* at 16+), whilst the non-academic pupils (those expected to complete full-time schooling at school-leaving age) attend the *Hauptschule* (a small percentage transferring to the *Realschule* at 15+).

school and five native English teachers. 46 native English pupils, 23 in a *grammar school* and 23 in a *secondary modern school*[1], also wrote answers to the letters of elicitation. Their letters were marked and graded by two of the English teachers. The performance of the pupils and markers was analysed in detail.

The test

There were two reasons for choosing a letter written in response to a letter of elicitation as a means of testing the pupils' communicative abilities in writing. Firstly, letter-writing is an objective in elementary English teaching in all three types of school and so the pupils could be expected to have been adequately prepared to cope with the task. Secondly, the task is one which validly reflects the characteristics of real communication: it calls for the ability to *interact* with a partner and to *use* language rather than display knowledge about its elements or systems; the language is used for a *purpose*; the stimulus is *authentic* and the language produced *open-ended*; both are situated in a real *context* with implications for appropriacy and discourse skills.

The letter of elicitation took the form of a letter from an English pupil to a German penfriend commenting on an earlier letter from the penfriend, and making enquiries and arrangements for a forthcoming summer holiday visit to England by the German penfriend. It was designed to elicit, or elicit a response to, certain specific speech functions (such as making comparisons, giving information, responding to an invitation) taken from the curricula of the different school types. There was in fact a different letter for each school type, though those for the *Gymnasium* and *Realschule* were exactly the same in content and different only marginally in the complexity of language.

The three letters were piloted with groups of pupils from the appropriate school types to ensure that they elicited the expected performance, after which some minor adjustments were made to the wording to

---

1 In the English selective system (now largely replaced by comprehensive schools), the *grammar school* can be roughly equated with the *Gymnasium/Realschule* and the *secondary modern school* with the *Hauptschule*.

produce final versions. (The letters are reproduced as Appendix A.)

The definitive letters were given to groups of 20 pupils each in a *Gymnasium*, *Realschule* and *Hauptschule*. The pupils were in the ninth grade (aged 15+) and fifth year of learning English. (A letter written by a pupil of the *Realschule* is reproduced as Appendix B.) The *Gymnasium* letter was also given to 23 pupils of an English *grammar school* and the *Hauptschule* letter to 23 pupils of an English *secondary modern school*. These pupils were in the fourth year of secondary schooling and also aged 15+.

There was a double purpose in giving the same test to English pupils. Firstly, we wished to confirm the authenticity of the task by seeing if native speakers of English responded to the letters in an essentially similar way to foreign learners. The results demonstrated unequivocally that they did. Secondly, we believed that the yardstick by which the performance of German pupils should be measured was not that of a perfect adult native speaker of English - an idealised concept - but that of a real native speaker of comparable age and academic ability.

The marking

Before submission to the markers the pupil letters were typed out to eliminate any possible influence that handwriting or layout might have but taking great care to preserve the exact wording, spelling and punctuation of the originals.

Each set of German pupil letters was marked by three experienced teachers from the appropriate type of school who, however, (with the exception of one teacher from the *Hauptschule*) had had no contact with the pupils. They were also marked by five native English teachers, only one of whom was a foreign language teacher (of German and French), three being teachers of English and one a teacher of physics.

The German markers were asked to mark the set of pupil letters as a normal test (*Klassenarbeit*) indicating errors and awarding a grade on the usual 1 - 6 scale (1 - very good, 2 - good, 3 - very satisfactory, 4 - satisfactory // 5 - borderline, 6 - unsatisfactory; grades 1 - 4 are pass grades). They were asked to indicate linguistic errors as follows: .... spelling or punctuation error, ~~~~ slight error, ________ medium error, ════════ grave error. It was left entirely to their judgment which errors were to be treated as slight, medium or grave. Communicative errors were to be indicated as "style" for inappropriacy, "meaning?" for unclear and "meaning??" for incomprehensibility.

The same instructions were given to the English markers together with an explanation of the German grading system and a request for a verbal comment at the end of each letter on the pupil's positive achievement (range of expression, cohesion, etc.).

The letters written by the English pupils were marked and graded in exactly the same way as those of the German pupils but by two markers only, both of them English.

The analysis of results

Lists were drawn up of the grades given by all the markers and the means, standard deviations and correlation matrices were computed.

The errors identified by the markers were recorded on *marking survey charts* (one for each pupil) together with the degree of gravity assigned to the errors. This meant taking every error that any individual marker had identified and checking whether any of the other seven markers had also identified it and, if so, at what level of severity. Once this considerable task had been completed, the *marking survey charts* provided the point of departure for a whole series of analyses of both pupil and marker performance.

An important early need was to establish a criterion for deciding what was and what was not an error. Unanimity of the eight markers was the exception rather than the rule, if only because even the sharpest-eyed marker will nod. We decided therefore to look for majority agreement and, since we expected differences to emerge between the German markers (GM) and the English markers (EM), to separate the two groups of markers. This resulted in eight possible *error types*:

Type 1 Errors identified by all or a majority of both EM and GM

Type 2 Errors identified by all or a majority of EM and a minority of GM.

Type 3 Errors identified by all or a majority of EM and no GM.

Type 4 Errors identified by all or a majority of GM and no EM.

Type 5 Errors identified by all or a majority of GM and a minority of EM.

Type 6 Errors identified by a minority of GM and no EM.

Type 7 Errors identified by a minority of EM and no GM.

Type 8 Errors identified by a minority of both GM and EM.

In subsequent analysis we treated Types 1 - 3 as *real errors*, since in each case they had been identified by all or a majority of the English markers, and Types 4 - 8 as *pseudo errors*. Of the latter, only Types 4 and 5, i.e. those identified by all or a majority of the German markers, were considered further. Types 6 - 8, where there was never more than a minority of either group of markers in agreement, were regarded as *idiosyncratic errors*. The following table shows the results for the three types of school:[1]

Table 1. Survey of error types.

		GYM	RS	HS	ALL
Real errors	Type 1	203	260	433	896
	Type 2	19	43	41	103
	Type 3	13	20	13	46
Pseudo errors	Type 4	33	28	17	78
	Type 5	46	59	23	128
Idiosyncratic errors	Type 6	112	85	73	270
	Type 7	202	240	258	700
	Type 8	78	76	68	222
Totals		706	811	926	2443

The following is a very brief account of the main analyses which were carried out on the *real errors*.

1. *Error cause*

   For about 95 % of the errors it was possible to advance a plausible explanation such as mother-tongue transfer, over-generalisation of

---

1 The letters produced by the *Gymnasium* (GYM) pupils were on the whole longer than those of the *Realschule* (RS), which were longer than those of the *Hauptschule* (HS) (means: GYM - 213, RS - 172, HS - 129). Other things being equal, a short letter should contain fewer errors than a long one. To make valid comparisons between the school types, it was therefore necessary to adjust the figures to a standard mean length. The mean length chosen for this purpose was that of the *Gymnasium*, and the "raw" figures for the other school types were adjusted accordingly.

a rule or inadequate learning. By far the largest category was mother-tongue transfer (43%) followed by inadequate learning (30%).

2. *Error category*

Errors were assigned to various sub-categories of grammar, vocabulary and style, or to spelling. The largest category was grammar with 54% of the errors (the most frequent being tense errors) followed by vocabulary with 24%, spelling with 15% and style with 7%.

3. *Error gravity*

The mean gravity of each error was computed separately for the German and English markers, counting a slight error as 1, a medium error as 2 and a grave error as 3. The category of each error for which there was a discrepancy of .5 or more between the two groups of markers was recorded as an indication of a differing degree of irritation caused by the error.

4. *Errors causing a breakdown of meaning*

Relatively few of the errors (about 10%) caused a real breakdown of meaning. Of those that did, the majority (62%) belonged to the category of vocabulary.

5. *Text errors*

Both the German and the English markers tended to concentrate their attention on errors commited within the boundaries of a sentence. Although some of the markers made global comments at the end of the letter, such as "jerky" or "abrupt", they generally overlooked - or at least discounted - errors of discourse, particularly errors of inter-textual relations (between the letter of elicitation and the pupils' answers). A survey of the pupil letters revealed some 300 text errors, only about 18% of which had been identified. Interestingly, when we subsequently asked 45 English pupils of the same age as the testees to mark three sample letters, they were much more sensitive to this category of error than either the German or the English teachers had been. (An example of an English pupil's marking can be found in Appendix B.)

An analysis was also made of *pseudo errors*, Type 4 only (i.e. those for which there was not even minority support from the English markers). The German markers had "created" errors about evenly between grammar and vocabulary/style. The grammar "errors" were predominantly (69%) in the area of tense.

Teacher performance

A number of hypotheses or expectations were established, which focused on three aspects of the performance of the teachers as markers:

1. the reliability of the grades they award;
2. the kind of errors they identify and the degree of gravity with which they regard them;
3. differences emerging between German and English markers.

The main hypotheses will now be considered in the light of the detailed analysis of results.

Hypothesis 1

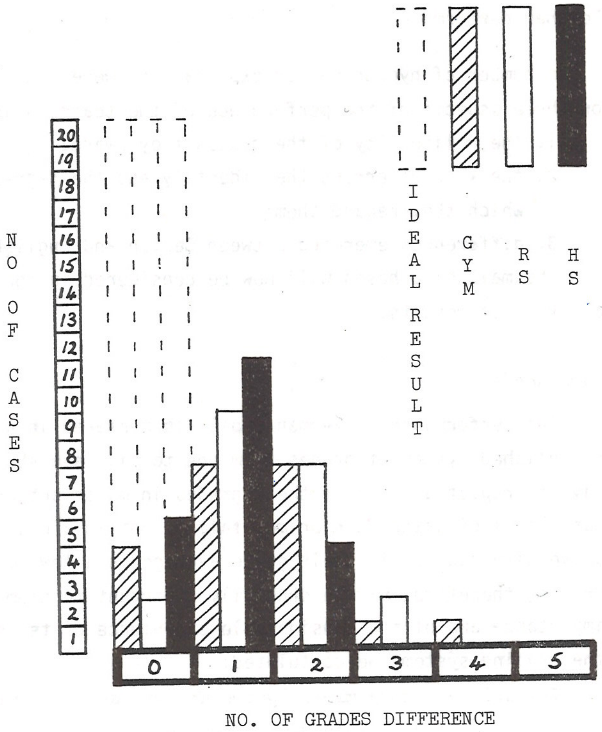
The performance of German pupils is measured in a uniform and well established system of grades from one to six (see above). Pupils may have to repeat a class if their grades in major subjects fall below the pass level of grade four or may fail to obtain a place on certain university courses if their grade average is a few decimal points too low. Grades, therefore, may be crucially important to them. In view of that importance and of the considerable experience of teachers in operating the grading system, we postulated:

The allocation of marks by the German teachers will display at the most a divergence of one grade.

This hypothesis could not be sustained for any of the school types, as the following table shows (Table 2.). The *Hauptschule* grades were the most consistent, meeting the hypothesis in sixteen cases out of twenty and never diverging by more than two grades, whereas the *Gymnasium* and *Realschule* both diverged by two or more grades in nine cases out of twenty. For one unfortunate pupil in the *Gymnasium* there was even a divergence of four grades!

It is conceivable that these grade differences disguise agreement between the markers on the relative standing of the pupils and result

Table 2. Grade differences between German markers



simply from the degree of leniency or severity with which the teachers convert their marks into grades. If that were the case, it would be revealed by high coefficients of correlation between the markers, and so the correlation matrices were computed for both the German and the English markers (whose grades diverged even more than those of the German markers). Table 3 shows the matrices together with the mean grade and standard deviation for each marker.

As can be seen, the correlations do not reveal any greater agreement between markers than the grade differences: in fact, they simply confirm the pattern, with the inter-marker reliability (German markers) for the *Hauptschule* still considerably higher than for the *Gymnasium* and *Realschule*.

The English teachers achieved somewhat better reliability for the *Gymnasium* and *Realschule* than the German teachers but much poorer reliability for the *Hauptschule*, so the overall picture for them too is of low inter-marker reliability.

We shall discuss possible reasons and remedies for this later, in the light of further hypotheses relating to the teachers as markers.

Hypothesis 2

The German markers' foreign language competence is of such high standard that on the whole they will not overlook very many errors.

If our criterion for what is a *real error* is accepted, then the German markers were certainly able to spot them on the whole, as inspection of Table 1 will verify. Type 1 errors (identified by all or a majority of both English and German markers) represent 86% of the *real errors* (Types 1 - 3). Moreover, there is no great difference between the German markers of the three different school types in this respect (GYM 86%, RS 80%, HS 89%).

Hypothesis 3

German markers will not identify as errors items that the majority of English markers accept as normal or suitable.

This hypothesis is refuted by the evidence. Error Types 4 and 5 are "errors" that German markers in effect invent, in that only a minority of English markers (Type 5) or no English markers (Type 4) agree that they

Table 3. Inter-marker correlations

mean
s.d.

German markers (GM) / English markers (EM)

Gymnasium

	GM1	GM2	GM3
GM1	4.1 1.0	.55	.37
GM2		3.7 1.2	.55
GM3			3.4 1.4

	EM1	EM2	EM3	EM4	EM5
EM1	3.2 1.0	.34	.52	.41	.56
EM2		3.6 0.8	.73	.81	.45
EM3			3.0 0.9	.78	.58
EM4				4.5 0.9	.71
EM5					2.4 0.7

Realschule

	GM1	GM2	GM3
GM1	4.1 0.9	.36	.33
GM2		3.5 1.3	.62
GM3			3.5 1.1

	EM1	EM2	EM3	EM4	EM5
EM1	4.7 1.0	.32	.47	.40	.58
EM2		4.4 0.6	.52	.62	.48
EM3			3.6 1.3	.66	.62
EM4				5.1 0.7	.58
EM5					2.9 1.1

Hauptschule

	GM1	GM2	GM3
GM1	3.8 1.0	.74	.79
GM2		3.2 1.1	.62
GM3			3.7 1.2

	EM1	EM2	EM3	EM4	EM5
EM1	5.7 0.7	.52	.48	.75	.59
EM2		4.5 0.8	.71	.51	.78
EM3			5.0 0.8	.55	.52
EM4				5.7 0.7	.38
EM5					3.9 1.2

are errors. As can be seen from Table 4, these *pseudo errors* form a significant proportion of the errors that all or a majority of German markers identify (Types 1, 4 and 5), in fact 19%. Here the picture is not uniform for the different school types: the teachers from the *Hauptschule*

Table 4. Distribution (%) of errors identified by all or a majority of GM

	GYM	RS	HS	ALL
Type 1	72	75	92	81
Type 4	12	8	3	7
Type 5	16	17	5	12

fare much better than their colleagues in the *Gymnasium* and *Realschule*. The teachers of the *Gymnasium* fare particularly badly, especially as regards Type 4 errors, for which there is no support at all from the English markers. However, it should be remembered that they are marking letters that are both longer and more sophisticated than those of the other school types, particularly the *Hauptschule*.

Hypothesis 4

The German markers will tend to overlook errors on the level of style and vocabulary.

Table 5 shows that the largest category of errors overlooked by the German markers (Types 2 and 3) was that of vocabulary and style, but grammar and spelling also accounted for large proportions of overlooked

Table 5. Distribution (%) of errors overlooked by GM

	GYM	RS	HS	ALL
Grammar	24	52	28	37
Vocabulary/style	58	34	57	48
Spelling	18	14	15	15

errors. It should be remembered that the actual number of such errors is small (see Table 1).

Hypothesis 5

The German markers will place more value on linguistic accuracy than the English markers, who will regard errors of vocabulary and style as particularly grave.

As described earlier (see *Error gravity*), we looked for cases where there was a clear discrepancy between the German and English markers in the severity with which they treated an error. We recorded the *error category* of all such cases. Table 6 shows the results.

Table 6. Error gravity discrepancy of $\geqslant 0.5$ between GM and EM

	GYM		RS		HS		ALL	
	GM	EM	GM	EM	GM	EM	GM	EM
Grammar	54	2	31	17	40	33	125	52
Vocabulary & style	8	10	22	6	13	64	43	80
Total	62	12	53	23	53	97	168	132

(Figures mean cases of greater severity.)

Overall, the figures support the hypothesis: the cases in which the German markers treat errors of grammar with greater severity than the English markers are more than twice as frequent as the reverse, whereas the English markers are almost twice as often more severe on errors of vocabulary and style. Overall, too, the German markers tend to be more severe than the English markers. However, the overall pattern is not repeated exactly for the individual school types.

Hypothesis 6

There is a type of semantic error which the German markers will overlook or underrate because they are able to interpret the English text through their knowledge of the pupils' mother tongue. Such errors will be regarded as grave by the English markers (with no knowledge of German) because they constitute a major impairment of meaning.

Errors did occur where an interpretation of the meaning of the English text was only possible through the mediation of German. "I pack it equal one" from a pupil of the *Hauptschule* is an extreme example. The context (a response to "Don't forget to bring your swimming things") and a word-for-word translation into German ("Ich packe es gleich ein") permit the interpretation "I'll pack it straight away". However, more often than not, the German markers treated such errors no less severely than the English markers. In only 25 cases out of 80 was the hypothesis upheld.

Hypothesis 7

Although a perfect specimen of marking is theoretically possible, no such specimen will occur in practice.

No single marker identified all of the *real errors* and so, in that sense, a perfect specimen of marking did not occur. In fact, it would be extremely unlikely to do so for two reasons. Firstly, for an open-ended task such as our test, no single perfect model of performance exists. In only a few cases was there full agreement between the English markers on what was an error. It was for that reason that we defined *real errors* as those errors identified by all *or* a majority of English markers. Two examples of usage where the English markers differed were the future tense (versus the present tense) and the ordering of time and place adverbs. Secondly, even the most conscientious and meticulous marker occasionally overlooks an error that every other marker identifies. In some cases this seemed to be attributable to a sort of fixation on a major error in the vicinity of the overlooked error.

Low inter-marker reliability - causes

There are five possible causes of the low inter-marker reliability that was revealed in the grading of the pupil letters:

1. the markers fail to agree on the errors in the letters;
2. the markers fail to agree on the gravity of the errors;

3. the markers have no common basis for evaluating the content of the letters;
4. the markers have no common basis for assessing the positive aspects of pupil performance, such as range of expression, length, independence from the language of the letter of elicitation;
5. the markers have no single system for converting error scores, positive marks, etc. into grades.

It is likely that a complex of these factors was at work in lowering the agreement between markers, but let us nevertheless examine them singly in the light of the evidence from the analysis of results.

1. *Disagreement over errors*

The survey of error types in Table 1 shows that the English markers agreed upon 1045 errors (Types 1 - 3) but disagreed upon an almost exactly equal number (1050 errors of Types 5, 7 and 8). The German markers agreed upon 1102 errors (Types 1, 4 and 5) and disagreed on a further 595 (Types 2, 6 and 8). The disagreement on what is an error, though much less for the German than for the English markers, is clearly substantial and must be a major contributor to the low reliability of both sets of grades.

2. *Disagreement over error gravity*

We have already seen in Table 6 that the German and English markers as a group often treated the same error with a different degree of severity. The same is true of individual markers: when they do agree on an error, then as often as not they disagree on its gravity. Table 7 shows the scale of disagreement for the German markers.

Table 7. Error gravity discrepancy between German markers

	GYM	HS	RS	ALL
Full agreement	107	126	110	343
Discrepancy of one degree of severity	79	54	110	243
Discrepancy of two or more degrees	14	24	43	81
Total discrepancy*	114	114	218	446

*In arriving at the "total discrepancy" figure, a discrepancy of two or more degrees of severity was weighted by a factor of 2.5, e.g. 114 = 79 + 14 x 2.5.

Clearly then disagreement over error gravity is also an important factor in the low inter-marker reliability.

3. *The evaluation of content*

There are two ways in which the content of the letters might be taken into account in assessment:

i) Does the pupil give the information asked for in the letter of elicitation?

ii) Is the letter rich in content or impoverished, interesting or dull, imaginative or pedestrian, etc.?

The first should clearly be considered in the marking as it is part of the communicative task. Comments by both German and English markers show that it was, by some of them at least. The second might be regarded as irrelevant to linguistic proficiency and possibly to communicative efficiency, but the comments which the English markers had been asked to make on positive aspects of the pupils' performance often refer to interest, originality, etc.

It is unclear, however, whether *all* the markers considered content and, if the did, *how* it was incorporated into their marking scheme. Content evaluation was probably a further source of poor reliability.

4. *The assessment of positive aspects of pupil performance*

All the English markers commented on positive aspects of pupil performance (it was part of their marking brief); some of the German markers did so occasionally. As with content, however, there was no evident basis for assessing such factors as complexity of structure and range of vocabulary, and this too no doubt played a part in reducing reliability.

5. *The system for converting marks into grades*

All the markers presumably had a system by which they arrived at a grade from their assessment of accuracy, content, expression, etc., but the system was rarely made explicit and is unlikely to have been the same for all the markers. Even had there been total agreement on the positive and negative aspects of the pupils' performance, different grades could still have resulted.

Low inter-marker reliability - remedies

Two preliminary attempts were made to improve the reliability of the marking. The first used an essay-marking technique proposed by John W. Oller, Jr. (Oller, pp. 385-91). The technique involves rewriting any portions of the essay that are clearly erroneous or unidiomatic in order to convey the scorer's best guess at the intended meaning. An "essay score" is then derived from the following formula:

ESSAY SCORE = the number of error-free words in the original essay minus the number of errors in the original essay divided by the total number of words in the corrected essay.

The technique was applied by one of the English markers in a re-mark of a sample of eighteen of the sixty German pupil letters, and the resulting essay scores were compared with the original grades. The trial raised a number of interesting questions about the technique, which will be reported on elsewhere (Maguire, forthcoming), but the resulting discriminatory power was insufficiently great, for these letters at any rate, for the essay scores to be readily translatable into the full range of the German grading system.

However, preliminary trials with a modified version of a technique proposed by D. Brodkey and R. Young in an article entitled "Composition Correctness Scores" proved more encouraging. We decided to embark on a full-scale trial of the technique by having our original German markers remark all the letters. As more than a year had elapsed since the first marking and the markers had no record of the errors they had identified or the grades they had awarded on that occasion, we felt confident that they would be able to operate the suggested system in a reasonably untrammelled fashion. The instructions the teachers received were as follows:

1. Underline all errors, including those of text-grammar (such as abrupt transition, omission of appropriate reference to letter of elicitation, lack of cohesion) as follows:

   ~~~~~~~~ - slight error
   ________ - medium error
   ======== - grave error

2. Guidelines for error gravity:

   ~~~~~~~~ - error impairing accuracy or appropriacy only slightly
   ________ - offence against basic rule of grammar, awkward choice of word(s), offence against text grammar/discourse organisation, omission of required information
   ======== - non-existent word, impairment of meaning, linguistic deviation apt to cause serious offence to native ear

N.B. Weight each error according to context, e.g. a grammatical error might be judged ________ in one context and ======== in another.
Penalise spelling errors only if glaring. Do not penalise punctuation errors.
Indicate missing information at end of letter by key word or phrase (e.g. "visit to school").

3. Count errors as follows:

   ~~~~~~~~ - 1   ________ - 2   ======== - 3

   and write the number in the margin. Add up the numbers to arrive at the "error score".

4. Divide the total number of words by the error score, correct to one decimal place. The result is the "correctness score".

5. Base the grades (1-6) on the correctness score. The level required for each grade is left to the teacher's discretion.

An example of a letter marked according to this procedure accompanied the instructions. A letter from one of the pilot groups of pupils was used for this purpose. To save the teachers effort and to ensure uniformity, the total number of words in each letter was marked on the copy sent to the teachers.

It will be noticed that the procedure the teachers were asked to operate was not very different from the standard type of marking adopted on the first occasion and only slightly more onerous. It was therefore a reasonable system to expect them to operate.

What contribution could it be expected to make to reducing or eliminating the causes of low inter-marker reliability outlined earlier?

Obviously, it can do little to improve agreement over what is an error, except that the teachers are now explicitly required to take account of errors of discourse, which, as we have seen earlier, they tended to overlook or discount, and puctuation errors are eliminated. Agreement over error gravity should be improved by the guidelines despite the element of
   ~~~~~~~~

subjectivity they contain. The evaluation of content is made explicit: missing information required by the letter of elicitation is to be penalised; no account is taken of interest of content. The system excludes the markers from incorporating positive aspects of pupil performance, such as range of expression, in the assessment except insofar as they must penalise serious lack of cohesiveness and offset error scores against length. No guidance is given on the conversion of correctness scores to grades.

Thus, the more explicit marking procedure makes some contribution at least to the reduction of four of the five areas of potential disagreement between markers. It provides a firmer basis for grading but still leaves the teacher to draw up the grade boundaries. We might expect therefore that the second marking would produce an improvement in the correlations between markers, which reveal how well the markers agree on the relative standing of the pupils, and probably but not necessarily, in the agreement on grades. That, as we shall see in the next section, is the kind of result we obtained.

The results of the second marking

The correlation matrices for the second marking are given in Table 8, alongside those for the first marking.

As can be seen, the agreement between the markers is greatly improved. Of the nine possible correlations, seven coefficients are higher - some of them considerably - one is unchanged and one is lower. All the correlations now reach significance, eight at the 1% level and one at the 5% level; whereas, on the first marking, only four correlations were significant at the 1% level, two reached the 5% level and three failed to reach significance. The *Gymnasium* and *Realschule* show large improvements in reliability, whereas the *Hauptschule* suffers a slight decline. There was, of course, far more room for improvement in the *Gymnasium* and *Realschule*.

The effect of the remarking exercise on grade discrepancy is shown in Table 9.

Table 8. Inter-marker correlations

GYMNASIUM

	GM2	GM3
GM1	.55* / .69**	.37 / .74**
GM2		.55* / .84**

REALSCHULE

	GM2	GM3
GM1	.36 / .76**	.33 / .50*
GM2		.62** / .75**

HAUPTSCHULE

	GM2	GM3
GM1	.74** / .74**	.79** / .68**
GM2		.62** / .67**

(SIGNIFICANCE: * = P < .05 ** = P < .01)

Table 9. Grade differences between German markers on first and second marking

No. of grades difference	GYM		RS		HS		ALL	
	1st	2nd	1st	2nd	1st	2nd	1st	2nd
0	4	5	2	4	5	2	11	11
1	7	11	9	12	11	9	27	32
2	7	4	7	4	4	6	18	14
3	1	0	2	0	0	3	3	3
4	1	0	0	0	0	0	1	0
	No. of pupils							

Overall, there is an improvement in how well the markers agree on grades: Hypothesis 1 (*The allocation of marks by the German teachers will display at the most a divergence of one grade*) now holds good for 72% of the pupils compared with 63% on the first marking. However, whilst the *Gymnasium* and the *Realschule* both show a large improvement from 55% to 80%, there is a large decline for the *Hauptschule* from 80% to 55%.

Inspection of the individual markers' means shows that GM3 (German marker no. 3) for the *Hauptschule* with a mean grade of 3.3, is a whole grade less severe in grading than GM1, with a mean of 4.2, and GM2, with a mean of 4.5. If GM3's grades are all lowered by one (the one grade 6 remaining 6), then GM3's mean becomes 4.3 and the second marking column of Table 9 for the *Hauptschule* reads 6, 7, 7, 0, 0. 65% of the *Hauptschule* grades then meet Hypothesis 1, a less severe decline from the first marking but a decline nevertheless, and overall, Hypothesis 1 holds good for 75% of the pupils. The correlations for the *Hauptschule* are scarcely affected (.67 becomes .66, otherwise no change) by the adjustment of GM3's grades, demonstrating that reasonable agreement over the relative standing of the pupils may be concealed by the arbitrary pitching of grade levels.[1]

---

1 It was interesting to discover *after* analysing the results of the remarking, and attributing the decline in reliability in part at least to GM3, that GM3 had in fact left school-teaching about a year before the second marking and might therefore be considered somewhat out of touch.

Conclusion

Although most of the many interesting aspects of the performance of German and English pupils in the letter-writing task we set them are outside the scope of this article, it is relevant to record that all of the German pupils succeeded more or less well in understanding the letter of elicitation and supplying the information it called for comprehensibly, if in some cases rather primitively. The range of performance of the English pupils was very large, and there were interesting parallels between *Gymnasium* and *grammar school* pupils on the one hand, and *Hauptschule* and *secondary modern school* pupils on the other hand. They suggest that the appropriate target level for the German pupils should not be the performance of an idealised adult native speaker but the real communicative performance of English children of the same age and ability range.

A guide to that target level would be an important contribution to improving the very low inter-marker reliability that we found in the assessment of an otherwise valid communicative task. On the one hand, it would be relevant to the definition of what is to be considered an error and, on the other hand, it would provide a basis for determining grade levels - both factors in reliability. As we have seen, other sources of poor reliability can be much reduced by the application of a system that is neither very different from nor much more time-consuming than standard (German) marking procedure.

References

Oller, J.W. (1979) *Language Tests at School*. London: Longman.

Maguire, D. (forthcoming) Essay marking: a trial run of Oller's method compared with German traditional method. University of Munich.

Brodkey, D. and Young, R. (1981) Composition Correctness Scores, *TESOL quarterly*, vol. 15, no. 2, 159-67.

Appendix A

Letter of elicitation for *Gymnasium/Realschule*

Manchester, 12th May, 1981

Dear Birgit,

Thanks for your last letter, which you wrote just before you went to spend Easter with your aunt. Did you have a good time?

Aren't you lucky - going with your parents to Italy this summer! We were in Florence and Rome two years ago and had a lovely time. Are you going there? Perhaps I can give you some tips on where to eat and what to see.

Gymnasium /Your Italian trip doesn't mean you won't be coming to stay with us in England, does it?/ *Realschule* /You <u>are</u> still coming to stay with us in the summer, too, aren't you?/ I am <u>so</u> looking forward to showing you all the sights and introducing you to all my friends. Is there anything you would really like to do or anywhere you've always wanted to go? Do tell me and we'll try to arrange it. By the way, I shall still be at school for the first week of your stay. Do you want to come with me?

Talking of school, how is your new English teacher? Is he (or perhaps it's a she this time) as nice as the one I met last year?

Must close now. Do write soon, won't you, and tell me you are still coming. Bye for now.

Love,

Linda

Appendix A cont'd

Letter of elicitation for *Hauptschule*

Liverpool, 12th May, 1981

Dear Georg,

Thank you for your last letter. It's great that you can stay with us in the summer. When will you be able to come? Do you think you can find your way across London? If you are not sure, we will come to meet you. When will you be arriving?

On 15th August we are going to the seaside for a week in our caravan. You can come with us if you can stay that long. Do you know how to sail? If not, I can teach you. Don't forget to bring your swimming things.

Did I tell you I've got a Saturday job now? I earn eight pounds a week, so I buy a lot of new records. Have you bought any new ones lately?

How is Uschi? Is she still mad about horses? Give her my love and your parents, too.

Best wishes,

Yours,

David

Appendix B

Letter written by a German pupil (*Realschule*) and corrected by an English pupil (*comprehensive school*)

Hallo Linda,

Thanks for your nice letter and for your invitation.

I enjoyed the Easter holidays very much.

quite *funny* *is*
My aunt is still very young, she is humorous and a good friend of me.

went *and* *and we played* *went*
We were swimming, dancing, playing golf and riding. The weather was

Have you got *relatives*
very good! Do you have such a nice aunt, too?

I'm very lucky to go to Italy this year. We are not going to Florence

so that we can swim in the sea
or Rome. We are going to the coast in order to swim very much.

very interested in *of England* *would like to meet*
I'm so curious of the sights and your friends. How old are they?

would also enjoy coming
I would like to go to an English football-match. I want to come

a lot *would like*
with you to school. There is much I want to see in your school:

your teachers, your time-table and your books. You think, I'm silly

because I want to go to school with you during my holidays. But

would be *nice*
I think it will become very funny!

elderly *and is a very*
But now to my English teacher. He is an old, always surly person.

doing very well
This year I'm not very good in English. The teacher I had last year is

now in an other school in the next town. It's a pity!

Now
The weather is fine and I want to go for a walk.

Goodbye!

Sauli Takala
Kasvatustieteen tutkimuslaitos
Jyväskylän yliopisto

SOME PERSPECTIVES ON CRITERION-REFERENCED MEASUREMENT

Introduction

There are several reasons for the recent interest in a direction in measurement and evaluation which frequently is referred to as "criterion-referenced" measurement. By criterion-referenced measurement is usually meant a type of measurement that is deliberately constructed to yield scores that are directly interpretable in terms of specified performance standards related to specific classes or domains of tasks (Glaser, 1963; Glaser and Nitko, 1971; Popham, 1978).

Classical test theory, which formed the basis of the psychological study of individual differences, abilities, etc., has had only tenuous links with learning theory (Cronbach, 1957; Glaser and Nitko, 1971). It has not, therefore, been much interested in aptitude-treatment interaction. Serious practical work on truly individualized instruction on a scientifically sound basis is of relatively recent origin. When interest in adaptive instructional systems grew in the 1960's, it became evident that there was a need for tests that are very sensitive to the content of individualized programs (Glaser, 1963).

Another reason why criterion-referenced measurement became a topic of growing interest is that when increasingly large sums of money became available through national budgets for experimental educational programs, it became a standard practice to require a research-based evaluation of the effectiveness of the programs. Since the contents of such programs were often based on new ideas about content and treatment, it was to be expected that standardized tests would not be considered very suitable to measure the effects obtained. More program-specific tests were needed.

A third reason for increased interest in criterion-referenced measurement derives from the growing demands for proof that national educational systems are working in a satisfactory way and that the money allocated to cover educational costs is well spent. When the performance of national systems are assessed, there is a growing interest to make sure that tests measure what has been taught and that test results tell the general public what students can do and what they cannot do.

A fourth reason is more closely related to decisions concerning individual students. When almost automatic promotion from grade to grade has become a pattern with the introduction of comprehensive-type educational systems, there has been growing concern that students may be promoted without having learned the knowledge and skills needed in the subsequent grade ("social promotion"). If school systems decide to adopt a stricter promotion policy, it is important that the amount of the risk of making false decisions is minimized. Program-specific tests are a useful tool in administering such a promotion policy.

This paper will first review some major sources of criterion-referenced measurement and describe briefly some alternative conceptualizations. Some comparisons are made between criterion-referenced and norm-referenced measurement. After that, stages in CRM are described with major emphasis on methods of content specification and the construction and selection of items. The paper then moves to discuss standard setting as an issue in CRM. This is followed by a review of how validity and reliability are treated in CRM. The paper concludes with a brief account of the uses of CRM and of current problems and issues in CRM.

1. Criterion- and Norm-Referenced Measurement

It was estimated that there were some 600 references on criterion-referenced measurement towards the end of the 1970's. Practically all of them were published during that decade. Yet, criterion-referenced measurement is not such a new idea.

E.L. Thorndike wrote about the difference between absolute and relative measurement some seventy years ago. Around 1950 Vahervuo in Finland carried out several studies on absolute and relative grading and on their theoretical basis. Still, it was in an article by Robert Glaser in 1963 that the term "criterion-referenced test" was introduced. The idea was favorably received but it did not lead to further work until in 1969 when Popham and Husek took up the concept and explicated further some of its implications.

Programmed learning and the behavioral objectives movement (e.g. Mager, 1962) were a major source in the emergence of criterion-referenced measurement. Carefully outlined teaching programs will not lead to a normal distribution of scores if the programs are, indeed, effective. There should be a high percentage of high scores and a decrease in variance. The latter is problematic for classical test theory, because most of its indices rely heavily on variance. Thus, it seemed necessary to conclude that variance-based estimates of test reliability are less appropriate in mastery-type instructional programs since they would unjustifiably label criterion-referenced tests as being of low reliability. New approaches were clearly needed (Popham and Husek, 1969).

Another major source, which is related to programmed learning and individualized learning programs, is the work done to discover learning hierarchies and curriculum (task) hierarchies (Gagne et al, 1962; Resnick, 1967). This work revealed that the testing of learning outcomes requires a thorough analysis of the subject matter as a preliminary step to item construction.

Criterion-referenced testing has been defined in a number of ways. According to Berk (1980), at least fifty different definitions have been proposed since Glaser's first paper. Perhaps the most concise definition has been suggested by Popham (1978, p. 93): "A criterion-referenced test is used to ascertain an individual's status with respect to a well-defined behavioral domain." This means that the interpretability of the test result is of primary concern. Whereas in norm-referenced measurement an individual's test score derives its meaning mainly from its relationship to the scores of other examinees (relative interpretation), the scores on a criterion-referenced test derive their meaning from the scores' relationship to a class or domain of tasks (absolute interpretation). Thus a domain score can be interpreted in terms of what an individual can do and what he cannot do and it also indicates what proportion of all possible tasks (items) of the whole item universe the individual could have solved if they were administered to him rather than only a sample of them. A domain score lends itself to absolute interpretations and can be used both for qualitative and quantitative descriptions (what is mastered and how much is mastered).

Several terms for this kind of testing have been proposed within the criterion-referenced movement. Ebel (1962) proposed a term "content-standard test" to describe a test which produces test scores which indicate what percentage of a systematic sample of defined tasks a person has solved correctly. Osburn (1968) used the term "universe-defined test" to refer to a test which produces an unbiased estimate of his score in an explicitly defined item content universe. Hively (1962) prefers the term "domain-referenced test" as a less ambitious term than universe-defined test. Carver (1974) has advocated the use of edumetric (rather than traditional psychometric) tests to measure within-individual growth (competence) instead of between-individual differences (ability, intelligence).

The term "objectives-based test" has sometimes been used as a near-synonym for criterion-referenced tests. If the items are simply derived from behavioral objectives without a strictly predetermined procedure, however, objective-based tests do not lend themselves to criterion-referenced interpretation.

The term "mastery test" has been derived mainly from the mastery learning system developed by Bloom (1968, 1971), largely on the basis of the model of school learning proposed by Carroll (1963). The main purpose of mastery tests is to help in the classification of students as masters or nonmasters of an objective in order to facilitate the management of an individualized teaching program.

If one were shown a test which only contained the instructions to students and the test items, it would be difficult to say whether the test is a criterion-referenced test or a norm-referenced test. In order to be able to make that decision it is necessary to know how the test was produced. It is in the work prior to the assembly of a test that most of the effort needs to be spent in producing a criterion-referenced test. Differences between two forms of criterion-referenced testing (domain-referenced and mastery tests) and norm-referenced testing are summarized in Table 1. The first five stages in the development of tests refer to the planning stage and the rest to the technical aspects of tests and their uses.

TABLE 1. Characteristics of Two Types of Criterion-Referenced Tests and of Norm-Referenced Tests (adapted from Millman, 1974, and Berk, 1980).

Stages of Development	Alternative Conceptualizations		
	Criterion-Referenced Testing		Norm-Referenced Testing
	Domain-Referenced	Mastery	
1. Specification of Content Domain	Maximum specification of content limits Methods: 1. Item transformations 2. Mapping sentences 3. Algorithms 4. Item forms 5. Amplified objectives 6. Test specifications	Content limits only partially specified Methods: Instructional and behavioral objectives	Content limits only partially specified Methods: Instructional and behavioral objectives
2. Item Construction	Generation rules	Traditional rules	Traditional rules
3. Specification of Item Domain	Infinite or finite item universe	Infinite ?	Infinite ?
4. Item Analysis	Purpose to detect flawed items Methods: 1. A priori judgement of item-objective congruence by subject matter experts 2. A posteriori computation of item statistics	Purpose to detect flawed items Methods: ?	Purpose to select items Methods: A posteriori computation of item statistics
5. Item Selection from Item Universe	Random	Nonrandom (?)	Nonrandom

Table 1 (cont.).

Stages of Development	Alternative Conceptualizations		
	Criterion-Referenced Testing		Norm-Referenced Testing
	Domain-Referenced	Mastery	
6. Cut-off Score Selection	Optional	Required	Required (?)
7. Validity	Content Construct Decision	Content Criterion-related Construct Decision	Criterion-related
8. Reliability	1) Consistency of decisions ($\hat{p}_o$, $\hat{k}$) 2) Dependability ($\Phi(\lambda)$) 3) Error of measurement or estimate around domain score	Consistency of decisions ($\hat{p}_o$, $\hat{k}$)	Traditional procedures (based on correlation)
9. Score Interpretation	Performance in relation to domain (level of functioning) Performance in relation to required level of mastery	Performance in relation to required level of mastery	Performance in relation to other examinees
10. Item and Test Variance	Not required	Not required	Required

2. Stages in Test Construction

2.1. Specification of Content

It is in the specification of the content domain that the greatest challenge and also the greatest merit of criterion-referenced testing lies. In traditional norm-referenced tests the content limits are only partially specified. Short instructional and behavioral objectives are used as the basis for item generation. As Bormuth (1970) and Anderson (1972), among others, have shown, there is so much room left for interpretation that the items may reflect the characteristics of the test constructor more than those of the instructional program. Too much room is left for creativity, which according to Popham (1978, 1980), is not as desirable as strict adherence to the content limits. Several methods have been proposed for making domain specification more adequate. These will be discussed below in some detail, since this is a crucial part of all criterion-referenced measurement.

Item Transformations

Bormuth (1970) has suggested that linguistic analysis based on transformational grammar could be used to make explicit the methods by which items are derived from statements of instructional objectives. Bormuth advocates operationalism as a way of introducing rigor into item construction and sees syntactic operations as a promising way to do this. His method is illustrated below. It shows some item transformations that have been performed on a sentence "The older sister put out the fire." Using syntactic transformations several comprehension questions could be asked about the sentence.

It seems obvious that Bormuth's method is a useful tool for generating items testing the comprehension of written and spoken discourse. Anderson (1972) provides some other examples of ways of generating questions to test discourse comprehension. One weakness of these methods is, however, that the emphasis is on sentence level operations rather than discourse level units. Recent work on discourse analysis by Halliday and Hasan, van Dijk, Meyer and others will be of use in moving from sentence to discourse-level testing.

Transformation Name	Question
Echo	The older sister put out the fire?
Tag	The older sister put out the fire, didn't she?
Yes-No	Did the older sister put out the fire?
Noun deletion	Who put out the fire?
	What did the older sister put out?
Noun modifier deletion	Which sister put out the fire?

Using these examples of item transformation, supply answers for Problem Set 2.

Problem Set 2

Item Transformations

The following statement appears as part of a paragraph in a science unit on balance scales: The heavier object is closer to the ground. Only items formed by the "yes-no" and "noun modifier deletion" transformations are to be used in a test to measure comprehension of this statement. What questions can be used?

1. Yes-No: ______________________

2. Noun modifier deletion: ______________________

Answers:

1. Is the heavier object closer to the ground?
2. Which object is closer to the ground?

(Source: Millman, 1974)

Mapping Sentence

Mapping sentences are used in facet analysis developed by Guttman (1969). Facet analysis can be used to describe the boundaries and structure of a domain of testing conditions. Facets are those dimensions or characteristics on which items in a given domain can differ. Facet analysis was used by the present writer in 1980 in an attempt to conceptualize the domain of written composition for the IEA International Study of Written Composition. The first attempt is illustrated below. (For a later version, see Takala, 1982).

Millman (1978) also used facet analysis in his study of how the form and content of items are related to item difficulty.

Mapping Sentence for the Domain of Writing
Following Guttman's Facet Analysis Scheme

A. Activity
1. Receive
2. Send

a/an

B. Channel
1. auditive
2. visual

message which deals with

C. Content/topic
1. self
2. school
3. home town
4. hobbies
5.
6.

and whose

D. Communication Partner
1. addressor
2. addressee

has/is

E. Role relationship between addressor and addressee
1. a higher social status
2. an equal social status
3. a lower social status
4. identical with addressor

and which is

F. Degree of publicity/formality
1. private
2. semi-public
3. public

consisting of

G. Input-output relationship (stimulus-response)
1. repetition of input
2. modification of input
3. internal input

and whose purpose is

H. Function
1. to preserve the message (documentative)
2. to inform (referential)
3. to persuade (emotive)
4. to describe (descriptive)
5.
6.

Different configurations of variables lead to different rhetorical modes (narrative, exposition, argumentation, etc.)

Examples:

A2 + B2 + C2 + D2 + E1 + F3 + G2 + H1 = a personal letter to a friend
A2 + B2 + C2 + D1 + E3 + F2 + G4 + H2 = a letter of application

Algorithms

The use of algorithms is closely related to facet analysis and mapping sentences. It also uses listing technique. The following example from Millman (1980) illustrates the use of algorithms in generating items.

The item generating process follows rules to ask for the cube of an integer between 4 and 7 and to produce four possible answers. The command MULTCHOICE letters the alternatives and randomizes them. The algorithm on line 40 generates the correct answer.

```
10 A=RANDOM (4,7)
20 A, SUPER (3), "EQUALS:"
30 FROM
40 A*A*A
50 A*3
60 A+3
70 A+30
80 A*10+3
90 A*A
100 (A=1)*A*A
110 MULTCHOICE
120 RIGHT 1
130 WRONG CHOOSE AT RANDOM
```

5^3 equals:

A. 53
B. 125
C. 8
D. 150

Answer: B

(Source: Millman, 1980)

Item Form

Perhaps the most sophisticated method of content specification is the so-called item form suggested by Hively (1968). Item forms serve two purposes (Hively et al, 1973): 1) they obviate the necessity to store individual items by substituting a set of written rules through which items can be generated when needed; and 2) they enable the relationship among items to be traced by giving clear specifications of relevant item characteristics. Thus there are two major parts in all item forms - one which tells how the items should be generated and another which describes their salient characteristics. One item form used by Hively and his colleagues in the Minnesota Mathematics and Science Teaching Project (MINNEMAST) is shown below.

ITEM FORM 16.14*

Comparing two objects on equal-arm balance and choosing a symbol to complete a statement of the weight relation.

GENERAL DESCRIPTION

The child is asked to compare the weights of two objects that may be (1) indistinguishable by hefting but easily distinguished on the balance, (2) indistinguishable even on the balance. In each of these situations, size varies as an irrelevant dimension. An equal-arm balance is available but instructions for its use are non-directive. The child is asked to select one of the three symbols (>, <, and =) and place it in the blank space provided between the two weight symbols.

STIMULUS AND RESPONSE CHARACTERISTICS

Constant for All Cells

The equal-arm balance is of similar construction to that used in MINNEMAST Unit 16, made of Tinkertoys, cardboard, string, a metal weight, and a foot ruler.

The objects are opaque, cylindrical bottles, identical except for weight (either 23 gm. or 25 gm.) and size (either 2" × ¾" or 2½" × 1¾"). Each is identified by a lower-case letter assigned at random.

The child is asked to complete a symbolic statement, corresponding to the weight relation, by choosing the correct relation symbol.

Distinguishing among Cells

Three weight relations (detectable by balance only, not by hefting or "feel") defined in terms of the location of the objects when placed in front of the child:

left > right; left < right; left = right.

Three size relations:

left > right; left < right; left = right.

CELL MATRIX

Weight Relations (Detectable by Balance Only)

Size Relations	$W_l > W_r$	$W_l < W_r$	$W_l = W_r$
$S_l > S_r$	(1)	(4)	(7)
$S_l < S_r$	(2)	(5)	(8)
$S_l = S_r$	(3)	(6)	(9)

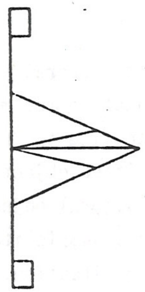
* Originally developed by Wells Hively.

ITEM FORM SHELL

MATERIALS	
Beam Balance Objects l and r from T.O. 16.14.0 Stimulus-Response sheet (attached) Pencil	
DIRECTIONS TO E Place materials in front of child (Keep order of objects given above.) Balance ← objects ← S-R sheet Subject	**SCRIPT** Here are two objects. They have symbols attached to them. Compare them by weight and write one of these three signs (point) in the blank (point) to form the comparison sentence. You may use this balance if you need to.

RECORDING

Attach Stimulus-Response sheet to this page.
Describe what child did.
If balance was used, insert object symbols in schematic drawing of the balance given below, and mark the position of the plumb-line at the time of child's judgment.



DESCRIPTION OF MATERIALS

Pencil (T.O. 16.1.1)

Beam Balance (T.O. 16.13.1): Equal-arm beam balance made from tinker-toy materials as decribed in MINNEMAST Unit 16.

Set of Weight Comparison Objects (T.O. 16.14.0): Set of opaque plastic cylindrical bottles with firmly fitting lids. Two sizes of bottles have been chosen. The small bottle has a length of 2" and a diameter of ¾". The large bottle has a length of 2½" and a diameter of 1¾". Two weight values have been chosen so that the objects cannot typically be distinguished by hefting but can be distinguished on the balance. Each object is designated by a randomly chosen, lower-case letter.

Size	Weight 23 gm	25 gm
small	a	m k
large	b	o n

Stimulus-Response sheet (attached to item) (T.O. 16.14.1), a sheet of paper approximately 6" × 4" with the following display:

Write >, <, or = in the blank

W_l ________ W_r

where l and r are the appropriate subscripts (from Replacement Scheme).

REPLACEMENT SCHEME

(l,r) Objects		
Cell 1:	(o,a)	
Cell 2:	(m,b)	
Cell 3:	Choose	from R.S. 16.13
Cell 4:	(b,m)	
Cell 5:	(a,o)	
Cell 6:	Choose	from R.S. 16.14
Cell 7:	Choose	from R.S. 16.15
Cell 8:	Choose	from R.S. 16.15
Cell 9:	Choose	from R.S. 16.17

REPLACEMENTS SETS

R.S. 16.13	Ordered pairs	(m,a);	(o,b)
R.S. 16.14	Ordered pairs	(a,m);	(b,o)
R.S. 16.15	Ordered pairs	(b,a);	(o,m)
R.S. 16.16	Ordered pairs	(a,b);	(m,o)
R.S. 16.17	Ordered pairs	(m,k);	(o,n)

SCORING SPECIFICATIONS

A correct response is made by writing the correct symbol (>, <, or =) in the blank space to complete the comparison sentence. This should be > in Cells 1, 2, and 3; < in Cells 4, 5, and 6; = in Cells 7, 8, and 9.

(Source: Hively et al, 1973)

As will be seen from the item form, any item form has the following characteristics (Osborn, 1968): 1) it generates items with a fixed syntactic structure, 2) it contains one or more variables (variable elements), 3) it defines a class of item sentences by specifying the replacement sets for the variables.

Such elaborate schemes as item forms guarantee that the domain is well defined and the population (universe) of items can be precisely described. It is, however, immediately obvious that to produce item forms must be very laborious and time consuming. It is also questionable whether similar levels of specificity can be reached in any other field than the formal languages of mathematics, logic and science.

Amplified Objectives

After finding out that item generation on the basis of traditional behavioral objectives was subject to too much interpretation and that using item forms was too demanding and led to "hyperspecificity", Popham (1980) worked with the so-called amplified objectives. As the name suggests, these are more detailed forms of behavioral objectives. They include 1) a brief statement of the objective, 2) a sample item, and 3) an amplified objective which specifies (a) the testing situation, (b) response alternative, and (c) criteria of correctness. The following example illustrates amplified objectives.

While amplified objectives clearly define the measured domain and specify item generation in greater detail than simple behavioral objectives, Popham (1980) observes that this attempt to "shoot for just the right balance between clarity and conciseness" failed. There was still too much room left for the personal interpretation of item writers.

Objective: **Given a sentence with a noun or verb omitted, the student will select from two alternatives the word which most specifically or concretely completes the sentence.**

Sample Item

Directions: Mark an "X" through one of the words in parentheses which makes the sentence describe a clearer picture.

Example: The racer (~~tumbled~~, went) down the hill.

Amplified Objective

Testing Situation

1. The student will be given simple sentences with the noun or verb omitted and will be asked to mark an "X" through the one word of a given pair of alternative words which more specifically or concretely completes the sentence.

2. Each test will omit nouns and verbs in approximately equal numbers.

3. Vocabulary will be familiar to a third- or fourth-grade pupil.

Response Alternatives

1. The student will be given pairs of nouns or pairs of verbs with distinctly varied degrees of descriptive power.

2. In pairs of verbs, one verb will either be a linking verb or an action verb descriptive of general action (e.g., is, goes), and one verb will be an action verb descriptive of the manner of movement involved (e.g., scrambled, skipped).

3. In pairs of nouns, one noun will be abstract or vague (e.g., man, thing), and one noun will be concrete or specific (e.g., carpenter, computer).

Criterion of Correctness

The correct answer will be an "X" marked through the more concrete, specific noun or through the more descriptive action verb in each given pair.

(Source: Millman, 1974)

Test Specifications

Experience with amplified objectives led Popham and his colleagues to believe that a so-called limited focus strategy was desirable. This means that the strategy is to focus measurement and to limit it to "a smaller number of assessed behaviors, but to conceptualize these behaviors so that they were large scale, important behaviors that subsumed lesser, en route behaviors" (Popham, 1980, p. 21).

The test specification consists of 1) a short general description, and 2) a sample item, which give the reader a general idea of what the test might contain. These are followed by 3) a detailed specification of the stimulus attributes and 4) response attributes including specification of the correct answer and, in the case of multiple choice items, of the reasons for various distractors. The test specification is illustrated below.

An Illustrative Set of Criterion-Referenced Test Specifications for a High School Minimum Competency Test in Reading

DETERMINING MAIN IDEAS

General Description

The student will be presented with a factual selection such as a newspaper or magazine article or a passage from a consumer guide or general-interest book. After reading that selection, the student will determine which one of four choices contains the best statement of the main idea of the selection. This statement will be entirely accurate as well as the most comprehensive of the choices given.

Sample Item

Directions. Read the selections in the boxes below. Answer the questions about their main ideas.

THE COLD FACTS

Had you lived in ancient Rome you might have relieved the symptoms of a common cold by sipping a broth made from soaking an onion in warm water. In Colonial America you might have relied on an herbal concoction made from sage, buckthorn, goldenseal, or bloodroot plants. In Grandma's time, lemon and honey was a favorite cold remedy, or in extreme cases, a hot toddy laced with rum. Today, if you don't have an old reliable remedy to fall back on, you might take one of thousands of drug preparations available without prescription. Some contain ingredients much like the folk medicines of the past; others are made with complex chemical creations. Old or new, simple or complex, many of these products will relieve some cold symptoms, such as a stopped-up nose or a hacking cough. But not a single one of them will prevent, cure, or even shorten the course of the common cold.

Reproduced with permission from *Test Specifications, IOX Basic Skill Tests: Secondary Level, Reading* (Los Angeles: The Instructional Objectives Exchange, 1978), pp. 21-24.

1. Which one of the following is the best statement of the main idea of the article you just read?
   a. Old-fashioned herbal remedies are more effective than modern medicines.
   b. There are many kinds of relief, but no real cures, for the common cold.
   c. Some of today's cold preparations contain ingredients much like those found in folk remedies of the past.
   d. Americans spend millions of dollars a year on cold remedies.

Stimulus Attributes

1. Each item will consist of a reading selection followed by the question "Which one of the following is the best statement of the main idea of the (article selection) you just read?" Eligible reading selections include adaptations of passages from factual texts such as general-interest books and consumer guides and pamphlets. Care should be taken to pick selections of particular interest to young adults and to avoid selections which may in the near future appear dated. Each reading selection will be titled, will be at least one paragraph long, and will contain from 125-250 words. Not more than 1,000 words of reading material can be tested in any set of five items. At least two of the five items in any set of five items must contain reading selections that are more than one paragraph long.

2. If necessary, the following modifications may be made to a selection used for testing:
   a. A title may be added if the selection does not have one, or if the selection represents a section of a longer piece whose title would not be applicable to the excerpt. If a title is added, it should be composed of a brief, interest-getting and/or summarizing group of words.
   b. A selection may be shortened, but only if the segment which is to be used for testing makes sense and stands as a complete unit of thought without the parts which have been omitted. If necessary, minor editing can be done to a reading selection which represents a shortening of a longer piece, but this editing should be for the purposes of clarity and continuity only, and not for the purposes of increasing or decreasing the difficulty level, or changing the content, of the text.
3. Reading selections used for testing should not exceed a 9th grade reading level, as judged by the Fry readability formula.

Response Attributes

1. A set of four single-sentence response alternatives will follow each reading selection and its accompanying question. All of these statements must plausibly relate to the content of the reading selection, either by reiterating or paraphrasing portions of that selection or by building upon a word or idea contained in the selection.
2. The three incorrect response alternatives will each be based upon a lack of one of the two characteristics needed by a correct main idea statement: *accuracy* and *appropriate scope*. A correct main idea statement must be accurate in that everything it states can be verified in the text it describes. It must have appropriate scope in that it encompasses all of the most important points discussed in the text that it describes.
3. A distractor exemplifies a *lack of accuracy* when it does any one or more of three things:
   a. Makes a statement contradicted by information in the text.
   b. Makes a statement unsupported by information in the text. (Such a statement would be capable of verification or contradiction if the appropriate information were available.)
   c. Makes a statement incapable of verification or contradiction; that is, a statement of opinion. (Such statements include value judgments on the importance or worth of anything mentioned in the text.)
4. A distractor exemplifies a *lack of appropriate scope* when it does one of two things:
   a. Makes a statement that is too narrow in its scope. That is, the statement does not account for all of the important details contained in the text.
   b. Makes a statement that is too broad in its scope. That is, the statement is more general than it needs to be in order to account for all of the important details contained in the text.
5. The important points which must be included in a main idea statement are those details which are emphasized in the text by structural, semantical, and rhetorical means such as placement in a position of emphasis, repetition, synonymous rephrasing, and elaboration. Whether any given main idea statement contains all of the important points that it should is always debatable rather than indisputable. The nature of the question asked on this test, i.e., select the *best* main idea statement from among those given, attempts to account for this quality of relative rather than absolute correctness.
6. The distractors for any one item must include at least one statement that lacks accuracy and one statement that lacks appropriate scope. On a given test, between 10 and 20 percent of the distractors should be sentences taken directly from the text.

7. The correct answer for an item will be that statement which is both entirely accurate and of the most appropriate scope in relation to the other statements given. If a sentence in the text itself qualifies as the best main idea statement which can be formulated about the selection, that sentence may be reiterated as a response option. No more than 20 percent of the items on a given test may have as their correct answer a main idea statement which is a direct restatement of a sentence in the text.

(Source: Popham 1980)

Popham (1980, 1981) feels that test specifications like the one shown in the above constitute a reasonable balance between clarity and conciseness so that busy people like teachers might not be put off by extreme specificity. Test specifications can also contain a supplement, which can give additional guidance in how to select stimuli, how to phrase questions, and so on.

2.2. Construction and Selection of Items

In the construction of items certain general rules have been devised for producing traditional norm-referenced tests. Such advice is presented in a number of books which deal with testing and evaluation. Most of these rules are also applicable to criterion-referenced measurement. The only difference is that more stringent demands are set for the procedure in item generation. It is, for instance, very important to stick to the limits set for the stimulus and response characteristics. Convergent rather than divergent creativity is needed in item generation. Work carried out by Carroll (1968, 1976) is of interest in this respect even if it is not in the mainstream of criterion-referenced measurement. Roid and Haladyna (1980) also provide a useful review of recent advances in the item-writing technology, including computer-based methods (cf. also Millman 1980). They note that the major positive result of the increased attention to the process of item writing is the heightened concern for the logical congruence between instruction and testing.

Once the rules for domain definition and for item generation have been worked out, it is necessary to consider specific items. Unlike in norm-referenced testing, it is necessary in criterion-referenced testing to know what the universe of items is that represents the defined domain content.

This universe can be finite or infinite. As Millman (1973) points out, it is not necessary that the population of items actually exists. What is necessary, though, is that the domain is so well described that a high agreement can be reached about what items are and what are not members of the population.

Further, unlike in norm-referenced and mastery tests, it is necessary to draw a random sample from the universe of all possible items because only this procedure makes it possible to produce an estimate of the examinees' total domains scores. Random sampling of items is needed in order to make it possible to generalize into the whole domain tested. It is generally assumed that 10-20 items are needed to measure a given content domain.

3. Standard-setting as an Issue in Criterion-Referenced Measurement

Standard-setting has been a topic of great controversy within the criterion-referenced movement. The need to set standards for acceptable performance has been especially great in mastery-type instructional programs, in which it is assumed that a certain level of mastery is optimal for both cognitive and affective outcomes (e.g., Block, 1972). Therefore, it would be important to identify masters and non-masters without making too many wrong classifications. In competency-based promotion systems it is also equally important to avoid too many cases of wrong decisions ("false positives", i.e., pseudo-masters and "false negatives", i.e., pseudo-non-masters). The decision-maker has to specify a loss function, in other words, state the relative seriousness of either passing students who lack requisite knowledge and skills or holding back students who in fact should be passed.

Methods available for setting standards have been discussed in several articles (Hambleton, 1980; Hambleton and Eignor, 1978; Hambleton et al, 1978; Jaeger, 1976; Meskauskas, 1976; Millman, 1973) and critically reviewed by Glass (1978). A whole issue of the Journal of Educational Measurement (Vol. 15, No. 4, 1978) was devoted to this problem. Glass provided a critical overview and Scriven, Block, Popham and Hambleton tried to rebut his main thesis that standard setting methods, in spite of their seemingly objective procedures, are basically arbitrary.

It is, in fact, now generally accepted that setting passing scores is arbitrary in the sense that it is based on judgment, but the advocates of standard setting maintain that it is not arbitrary in the sense of "capricious" or "unjustifiable". They point out that human life is full of situations where informed judgment must be exercised and measurement should not be faulted too much if some of its procedures also must resort to this method. Thus, they claim, what is needed is not the abolishment of standard setting but the improvement of its procedures. Hambleton (1980) classifies them into three groups: judgmental, empirical and combination.

All judgmental methods require that data are collected from subject-matter experts and other qualified judges for setting standards. Individual items are carefully inspected to judge how a minimally competent person would perform on them. Methods proposed by Nedelsky (1954), Angoff (1971), Ebel (1972) and Jaeger (1978) differ on some points, and what is more disturbing, they can lead to quite different passing scores and rates. It has been shown, for instance, by Andrew and Hecht (1976) that when the same judges used both the Ebel and Nedelsky methods, the passing scores varied from 49% of all items to 68% and the passing rates varied from 50% of all examinees to 95%. Such variability is clearly too wide and indicates the need for further work on this problem.

Since empirical methods are seldom used alone, they will not be discussed in this paper. The combination method uses both judgmental and empirical data. In the "Borderline Group Method" the judges are first asked to think of a minimally acceptable performance on the measured content area. They are then asked to give a list of those students whose performance is so close to the borderline that it is difficult to classify them with confidence. The test is then administered and the median score for the borderline group is taken as the passing score.

In the "Contrasting Groups Method" the judges are first asked to determine in their minds the minimally acceptable performance level and then identify those students who can be classified clearly either as masters or non-masters. Empirical test data are then obtained for both groups and the point of intersection of the two score distributions is taken as the passing standard. The present author used this method in 1979 in the first national assessment of English as a foreign language in Finland in an attempt to study how teachers' judgments could be used in establishing a common core syllabus for English. The results have not yet been analysed.

4. Validity as an Issue in Criterion-Referenced Measurement

Criterion-referenced tests are more and more often used in monitoring individual progress through objectives-based instructional programs (formative testing), to diagnose learning problems (diagnostic testing), to evaluate educational and social programs (program evaluation), and to assess level of performance on certification and licensing examinations. The usefulness of such applications depends heavily on the validity of the procedures undertaken in such testing.

According to Hambleton (1980) validity considerations in criterion-referenced testing arise at three steps: 1) the selection of objectives (content domain), 2) the measurement of objectives (content domains) included in the criterion-referenced test, and 3) the uses of test scores.

Validity is a difficult topic in all measurement and criterion-referenced measurement is no exception. Terminology varies quite a lot so that different terms are used to designate the same characteristic and the same term is used to designate somewhat different things. There are also some fundamental confusions that have persisted for a long time.

As Cronbach (1971), Messick (1975) and Linn (1979) have pointed out, a major conceptual confusion arises from the fact that content validity is focused on test forms rather than test scores, on instruments rather than measurements. In Linn's words "questions of validity are questions for the soundness of the interpretations of a measure... Thus, it is the interpretation rather than the measure that is validated. Measurement results may have many interpretations which differ in their degree of validity and in the type of evidence required for the validation process" (Linn, 1979, p. 109). For this reason, Messick states that content coverage is an important consideration in test construction and interpretation but it does not itself provide validity. He would prefer the term "content relevance" or "content representativeness", since they do not really provide evidence for the validity of the interpretation of scores.

Popham (1978) uses the term "domain-selection validity" to refer to the question of how well the results obtained can be generalized to as many other domains as possible. It thus resembles "construct validity" to some extent, although the latter is a more theoretical concept. Since testing for many reasons ought to be limited to a minimum, it is important to measure such domains and use such techniques which permit maximum generalization across domains of content. Domain-selection validity can be assessed by asking experts to give judgements on the relevance of selected domains.

Popham (1978) proposes them term "descriptive validity" to indicate the representativeness of measured content. In traditional norm-referenced testing no quantitative indices are usually given to describe content representativeness (cf. Table 1). In criterion-referenced testing, judges can be used to assess to what extent items are congruent with the test specification. Hambleton (1980) provides some useful methods for doing this. In some areas, where it is possible to specify completely a pool of valid test items, the representativeness of items can be ensured by drawing a random sample from the item pool. This was the procedure adopted when the present author studied students' active and passive vocabulary of English in the Finnish comprehensive school in 1979.

Hambleton (1980) uses the term "decision validity" to refer to the decisions made on the basis of scores. Popham (1978) uses the term "functional validity" in much the same sense. Decision validity in criterion-referenced testing is often related to standard setting (minimum passing scores). Since that question is discussed elsewhere in this paper (section 3, p. [illegible]), it will not be dealt with further in this context. A good review of decision-consistency is in Subkoviak (1980). Hambleton and Eignor (1978) and Walker (1978) review and assess standards and guidelines for evaluating criterion-referenced tests and test manuals.

5. Reliability as an Issue in Criterion-Referenced Measurement

Traditional methods of estimating reliability in norm-referenced measurement are usually based on correlational analyses where variance is a key concept. Since there may be relatively little variation in the scores of criterion-referenced tests, correlation-based estimates may not be ideally suitable for the estimation of reliability.

As Berk (1980) has noted there are at least three major conceptualizations of criterion-referenced test reliability: 1) consistency of mastery-non-mastery decisions across repeated measures with one test form or parallel test forms, 2) consistency of squared deviations of individual scores from the cut-off scores across parallel or randomly parallel test forms, 3) consistency of individual scores across parallel or randomly parallel test forms.

Subkoviak (1980) gives a good survey of five methods of determining decision-consistency reliability. Usually only two statistics are used in this context: P_0, which indicates the proportion of individuals consistently classified as masters and non-masters across parallel test forms, and κ, which estimates the proportion of individuals consistently classified beyond that expected by chance. Thus, P_0 estimates the overall consistency whereas κ estimates consistency due to testing alone. The choice of the index has to be based on whether one wants an estimate of overall consistency of decisions for whatever reason or of the contribution of the test alone. In most cases, it is probably advisable to report both estimates.

Brennan (1980) reviews the generalizability theory approach to reliability, which builds on the work by Cronbach and his associates (1972). Generalizability theory is based on the analysis of variance model and focuses on the estimation of various variance components in different types of test x items designs. Generalizability theory allows for the existence of many types and sources of error and it does not require strictly parallel tests for reliability estimation. Only randomly parallel tests are required.

As in the case of the decision-consistency approach, there are two indices of reliability (or dependability): $\Phi(\lambda)$ provides an estimate of the dependability of mastery-non-mastery decisions based on the testing procedure (λ represents the cut-off score), and Φ the "general purpose" index that is independent of the cut-off score and which can be used to estimate individual domain scores (a major interest in the present writer's study of the size of students' active and passive vocabulary). $\Phi(\lambda)$ is related to the reliability of criterion-referenced <u>test scores</u> and Φ is associated with the reliability of domain <u>score estimates</u>. The former indicates how closely the scores for any examinee can be expected to agree, the latter the degree of agreement with chance agreement removed. Thus $\Phi(\lambda)$ characterizes the dependability of decisions, or estimates, based on the testing procedure. Its magnitude depends, in part, on chance agreement. The index Φ characterizes the contribution of the testing procedure to the dependability of decisions, over and above what can be expected on the basis of chance agreement (Brennan, 1980).

As in the case of the decision-consistency approach, it might be useful to give both estimates. Brennan (1980) also strongly recommends that variance components too should always be reported.

6. Uses of Criterion-Referenced Measurement

The several possible applications of criterion-referenced measurement are mainly due to the increased rigor and precision in the description of important subject-matter domains and of behavior related to them. Some of the most common uses of CRM are described below drawing mainly on Millman (1974) and Popham (1978).

CRM can be used in needs assessments, which help in setting educational priorities. Need can be defined as the difference between an expected and the present observed situations. The latter can best be ascertained by means of CRTs, which possess a high degree of content representativeness. It also follows that CRM can be used in individualized teaching programs to assess students' current status with respect to objectives.

One of the most promising uses of CRM is in the area of large-scale program evaluation. Since CRM puts such rigorous demands on the item-program congruence, it is ideally suited to reveal the effectiveness of instruction or the lack of it. CRM with random samples of items from well-defined content domains provides reliable estimates of students' domain scores and makes reliable and valid generalizations to whole teaching programs possible. It furnishes reliable qualitative and quantitative information on learning and thus CRM will be of great help in efforts to develop education.

Methods developed within the so-called modern test theory movement, which has worked out new methods for avoiding some of the problems and limitations of classical test theory, (for instance, latent trait theory, generalizability theory, Bayesian methods), make it possible to shorten testing time. This is possible by either using the instructors' earlier knowledge of students in the estimation of their level of functioning (Bayesian methods) or by applying multilevel testing procedures or both combined. In multilevel testing items of varying difficulty, carefully prepared from some well-defined domain of content, are divided into a few groups. Each student takes the group of intermediate difficulty and then goes to either an easier or more difficult set depending on how difficult the intermediate set was for him or her. Lord (1976) notes that testing time can be reduced to one half and the number of items needed can be dropped from 100 to 20. Multilevel testing also has the positive affective effect of not shocking students with too difficult items or boring them with too easy ones.

7. Current Problems and Issues in Criterion-Referenced Measurement

As Popham (1978) points out one of the greatest problems in the development of criterion-referenced measurement is its difficulty and laboriousness. Where to get the resources for activity which presupposes highly trained full-time measurement experts and takes a lot of time? Popham (1980) says that he is distressed that he is unable to teach people how to go about the conceptualization of tested domains. In his own words, "at no point in the test development process for criterion-referenced measures is it more apparent that we are employing art, rather than science, than when the general nature of the behavioral domain to be tested is initially conceptualized" (p. 26).

It seems to the present writer that Popham is overemphasizing the "artistic" aspect of domain conceptualization. It seems likely that the reason for the felt difficulty is mainly due to the lack of a theoretical grasp of the structure and nature of the tested subject matter. If there were a better theoretical conception of the content structure and of the cognitive structure of some school subject, surely domain specifications would not need to be so much "artistic endeavors of no small shakes" (Popham, 1980, p. 27). It is, however, hard to find persons who master both the theoretical structure of subject-matter and the structure of the cognitive processes involved in its learning and use. Usually one is an expert in only one of these two aspects. After several years of work in curriculum construction, curriculum evaluation and textbook writing the present writer is convinced that the state of art in subject-specific domain specification in several school subjects is very low and serious work in this area has hardly been started. There is an urgent need for developing the "psychologies" of specific school subjects if there is to be any real progress in curriculum construction, teaching and evaluation.

Another related problem is the codification of guidelines for the construction of criterion-referenced tests and for their use. This would also be of great help in the training of test constructors and test users.

In addition to such content-specific problems, there are a number of technical problems that need to be studied. These include methods of estimating the validity and reliability of different uses of criterion-referenced tests; the use of computers in generating test items; and the employment of new ideas of modern test theory in criterion-referenced measurement.

Criterion-referenced measurement is so new if it is compared with norm-referenced measurement, and similarly modern test theory is new in relation to classical test theory, that both have a number of "unsolved problems and problematic solutions", as Popham so aptly puts it. There is intensive work being done all over the world to produce less problematic solutions to such problems and there is no need to doubt that such solutions will be forthcoming.

8. Discussion

Criterion-referenced measurement and norm-referenced measurement share a number of features. As in several other fields, for instance, in curriculum construction, new approaches usually mean only new emphases. At first there is a tendency to exaggerate differences. It is possible that this is inevitable when a new idea is introduced. Karl Popper has suggested that certain dogmatism may have an important part to play in the development of science, because giving up an idea too soon may mean that its merits and weaknesses are not given a sufficient chance of showing themselves. A scientist should not be too ready to adopt a new idea or to abandon an old one without persisting in some seemingly dogmatic stance for some time for the sake of argument. We should know how to play the believing and doubting games in a balanced way.

Criterion-referenced measurement shows some characteristics of this initial dogmatism. At first it was categorically stated that CRM does not need such concepts as item and score variance; that empirical item analyses are not needed; that norm data should not be gathered; and that content validity is the most important aspect of CRM. It was soon admitted, however, that these claims were overstated. Item variance usually occurs and serves a useful purpose in CRM testing as well as in norm-referenced testing. Similarly, it was conceded that norm data are not embarrassing for CRM. On the contrary, they add useful information and can help to interpret how "good" is "good enough". A posteriori empirical item analyses complement a priori judgemental (rational-logical) item analysis and help to detect flawed items. And, finally, content validity is not the all-important consideration in CRM. While content representativeness is a necessary characteristic of CRM it does not guarantee the validity of interpretations based on CRT scores.

Criterion-referenced measurement has the special advantage that it provides an exact description of a person's performance level in an entire domain and not only in the presented items. Several requirements must be fulfilled before such an interpretation is possible. First, there has to be a detailed description of the measured domain. Second, there must be a detailed description of the instrument, which includes the specification of the stimulus and response parts and of the scoring system. Third, items must be generated that have a high item-objective congruence and which are also a representative random or stratified random sample from the item pool. If CRM is used for program evaluation there must also be a representative sample of students from the entire population. In the latter case it is advisable to use matrix sampling with several parallel test versions rotated in the class.

One of the greatest attractions of CRM for the present writer is its emphasis on the conceptualization of measured domains. This lends support to his personal claim, which goes back several years, that one of the greatest obstacles for the development of teaching is the lack of theoretically sound conceptualizations of the units and processes in learning a particular subject matter. He would, therefore, fully agree with the view recently put forward by Popham:

> When created by instructionally astute developers, a criterion-referenced test can lay out so lucidly a set of teachable skills that the test itself becomes a potent force for instructional improvement. Instead of being an afterthought for use at the close of instruction, a properly conceptualized criterion-referenced test can stimulate measurement-driven instructional enhancement. Test developers can literally create test items so that they agree with one or more instructionally powerful explanatory constructs which teachers can then employ during their lessons... . This sort of focused instructional enterprise is not teaching-to-the-test in the negative sense that one teaches toward a particular set of test items. Rather, this approach constitutes teaching-to-the-skill, a highly effective and thoroughly defensible instructional strategy" (Popham, 1981, pp. 106-107).

Thus it might be that "the testing tail wagging the teaching dog" may not be such a problem or the embarrassment it is often taken to be if the tail is fully compatible with the dog. The present writer's personal experience with curriculum construction and evaluation, and with the in-service education of teachers in Finland suggests that the most effective and fastest way to promote desirable changes in teaching is to make sure that testing and tests display the characteristics of desirable student performance. Tests are the most concrete ways of signaling to teachers and students what the desirable content and forms of learning are.

Focusing on testing may be more effective than focusing on curricula and teaching materials since testing has a more limited scope and it is, therefore, possible to produce very carefully constructed tests that are, in a sense, modules of teaching. Such tests can serve as examples for preparing units of teaching and for individual lessons. By concentrating on important aspects of the subject matter it is possible to produce such modules which can also serve as a stimulus for textbook writers. While individual units and modules do not constitute an entire syllabus, they are useful wholes as such and can serve as useful models. Practical experience shows that it is much more difficult to seek to conceptualize an entire curriculum with similar rigor and it is also a huge task to produce a textbook package with a similarly consistent approach. Thus testing may, indeed, be a sensible starting point and lead to improved curricula and textbooks. At the very least, the potential contribution of work done within testing and measurement to curriculum design and instruction should not be ignored.

REFERENCES

Anderson, C. 1972. How to Construct Achievement Tests to Assess Comprehension. Review of Educational Research, 42, 145-170.

Andrew, B.J., and J.T.A. Hecht. 1976. A Preliminary Investigation of Two Procedures for Setting Examination Standards. Educational and Psychological Measurement, 36, 45-50.

Angoff, W.H. 1971. Scales, Norms and Equivalent Scores. In R.L. Thorndike (Ed.) Educational Measurement. Washington, D.C.: American Council of Education.

Berk, R.A. 1980a. Introduction. In R.A. Berk (Ed.) Criterion-Referenced Measurement: The State of the Art. Baltimore and London: The Johns Hopkins University Press, 3-9.

---------. 1980b. Item Analysis. In R.A. Berk (Ed.) Criterion-Referenced Measurement: The State of the Art. Baltimore and London: The Johns Hopkins University Press, 49-79.

Block, J.H. 1972. Student Evaluation: Toward the Setting of Mastery Performance Standards. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Bloom, B.S. 1968. Learning for Mastery. Evaluation Comment, Vol. 1, No. 1.

Bloom, B.S., J.T. Hastings, G.F. Madaus (Eds.) 1971. Handbook on Formative and Summative Evaluation of Student Learning. New York: McGraw-Hill.

Bormuth, R. 1970. On the Theory of Achievement Test Items. Chicago: University of Chicago Press.

Brennan, R.L. 1980. Applications of Generalizability Theory. In R.A. Berk (Ed.) Criterion-Referenced Measurement: The State of the Art. Baltimore and London: The Johns Hopkins University Press, 186-232.

Carroll, J.B. 1963. A Model of School Learning. Teachers College Record, 64, 723-733.

------------. 1968. The Psychology of Second Language Testing. In A. Davies (Ed.) Language Testing Symposium: A Psycholinguistic Approach. London: Oxford University Press, 46-68.

------------. 1976. Psychometric Tests as Cognitive Tasks: A New "Structure of Intellect". In L.B. Resnick (Ed.) The Nature of Intelligence. New York: Lawrence Erlbaum, 27-56.

Carver, R.P. 1974. Two Dimensions of Tests: Psychometric and Edumetric. American Psychologist, 29, 512-518.

Cronbach, L.J. 1957. The Two Disciplines of Scientific Psychology. American Psychologist, 12, 671-684.

----------. 1971. Test Validation. In R.L. Thorndike (Ed.) Educational Measurement. Washington, D.C.: American Council of Education.

Cronbach, L.J., G.C. Gleser, H. Nanda, and N. Rajaratnam. 1972. The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles. New York: Wiley.

Ebel, R.L. 1971. Content Standard Test Scores. Educational and Psychological Measurement, 22, 15-25.

---------. 1972. Essentials of Educational Measurement. Englewood Cliffs, N.J.: Prentice-Hall.

Gagne, R.M., J.R. Mayor, H.L. Garstens, and N.E. Paradise. 1962. Factors in Acquiring Knowledge of a Mathematical Task. Psychological Monographs, Vol. 76, No. 526.

Glaser, R. 1963. Instructional Technology and the Measurement of Learning Outcomes: Some Questions. American Psychologist, 18, 519-521.

Glaser, R., and A. Nitko. 1971. Measurement in Learning and Instruction. In R.L. Thorndike (Ed.) Educational Measurement. Washington, D.C.: American Council of Education, 652-670.

Glass, G.V. 1978. Standards and Criteria. Journal of Educational Measurement, 15, 4, 237-261.

Guttman, L. 1969. Intergration of Test Design and Analysis. In Proceedings of the 1969 Invitational Conference on Testing Problems. Princeton, N.J.: Educational Testing Service.

Hambleton, R.K. 1980. Test Score Validity and Standard-Setting Methods. In R.A. Berk (Ed.) Criterion-Referenced Measurement: The State of the Art. Baltimore and London: The Johns Hopkins University Press, 80-123.

Hambleton, R.K., and D.R. Eignor. 1978. Guidelines for Evaluating Criterion-Referenced Tests and Test Manuals. Journal of Educational Measurement, 15, 4, 321-327.

Hambleton, R.K., H. Swaminathan, J. Algina, and D.B. Coulson. 1978. Criterion-Referenced Testing and Measurement: A Review of Technical Issues and Developments. Review of Educational Research, 48, 1, 1-47.

Hively, E., G. Maxwell, G. Rabehl, D. Sension, and S. Lundin. 1973. Domain-Referenced Curriculum Evaluation: A Technical Handbook and a Case Study from the Minnemast Project. CSE Monograph Series in Evaluation, No. 1. Los Angeles: Center for the Study of Evaluation, University of California.

Jaeger, R.M. 1976. Measurement Consequences of Selected Standard-Setting Models. Florida Journal of Educational Research, 18, 22-27.

-----------. 1978. A Proposal for Setting a Standard on the North Carolina High School Competency Test. Paper presented at the annual meeting of the North Carolina Association for Research in Education, Chapel Hill.

Linn, R.L. 1979. Issues of Validity in Measurement for Competency-Based Programs. In M.A. Bunda and J.R. Sanders (Eds.) Practices and Problems in Competency-Based Measurement. Washington, D.C.: National Council on Measurement in Education, 108-123.

Lord, F.M. 1976. Test Theory and the Public Interest. Proceedings of the 1976 ETS Invitational Conference. Princeton, N.J.: Educational Testing Service, 17-30.

Mager, R.F. 1962. Preparing Instructional Objectives. Palo Alto: Fearon Publishers.

Meskauskas, J.A. 1976. Evaluation Models for Criterion-Referenced Testing: Views Regarding Mastery and Standard-Setting. Review of Educational Research, 46, 133-158.

Messick, S.A. 1975. The Standard Problem: Meaning and Values in Measurement and Evaluation. American Psychologist, 30, 955-966.

Millman, J. 1973. Passing Scores and Test Lengths for Domain-Referenced Tests. Review of Educational Research, 43, 205-216.

----------. 1974. Criterion-Referenced Measurement. In W.J. Popham (Ed.) Evaluation in Education: Current Applications. Berkeley: McCutchan.

----------. 1978. Determinants of Item Difficulty: A Preliminary Investigation. Center for the Study of Evaluation, CSE Report No. 114.

----------. 1980. Computer-Based Item Generation. In R.A. Berk (ed.) Criterion-Referenced Measurement: The State of the Art. Baltimore and London: The Johns Hopkins University Press, 32-43.

Nedelsky, L. 1954. Absolute Grading Standards for Objective Tests. Educational and Psychological Measurement, 14, 3-19.

Osborn, H.G. 1968. Item Sampling for Achievement Testing. Educational and Psychological Measurement, 28, 95-104.

Popoham, W.J. 1978. Criterion-Referenced Measurement. Englewood Cliffs, N.J.: Prentice Hall.

------------. 1980. Domain Specification Strategies. In R.A. Berk (ed.) Criterion-Referenced Measurement: The State of the Art. Baltimore and London: The Johns Hopkins University Press, 15-31.

------------. 1981. Measurement Essentials for the Essentials of Education. In L.Y. Mercier (Ed.) The Essentials Approach: Rethinking the Curriculum for the 80's. U.S. Department of Education, 97-115.

Popham, W.J., and T.R. Husek. 1969. Implications of Criterion-Referenced Measurement. Journal of Educational Measurement, 6, 1-9.

Resnick, L.B. 1967. Design of an Early Learning Curriculum. Working Paper 16. University of Pittsburgh: Learning Research and Development Center.

Roid, G., and T. Haladyna. 1980. The Emergence of an Item-Writing Technology. Review of Educational Research, 50, 2, 293-314.

Subkoviak, M.J. 1980. Decision-Consistency Approaches. In R.A. Berk (Ed.) Criterion-Referenced Measurement: The State of the Art. Baltimore and London: The Johns Hopkins University Press, 129-185.

Takala, S. 1982. On the Origins, Communicative Parameters and Processes of Writing. Evaluation in Education, 5, 3, 209-230.

Walker, C.B. 1978. Standards for Evaluating Criterion-Referenced Tests. Center for the Study of Evaluation, CSE Report No. 103.

Edgar Süßmilch
Universität Düsseldorf
Düsseldorf, BRD

SPRACHLEISTUNGSMESSUNG MITTELS C-TESTS*

1. Einleitung

Im folgenden Beitrag wird ein Überblick gegeben über die Entwicklung eines neuartigen Testverfahrens zur Messung der allgemeinen Sprachbeherrschung. Dieses Testverfahren, 'C-Test' genannt, ist aus der kritischen Diskussion heraus um die theoretischen Grundlagen und die angestrebten Ziele einer Diagnostik im Sprachunterricht entstanden. Der C-Test ist insbesondere eine Antwort auf die Kritik an einem anderen globalen Sprachtest, dem Cloze-Verfahren. Auf die angesprochenen Aspekte, vor allem aber auf die empirische Prüfung der Anwendung von C-Tests, die eine Modifikation des Cloze-Verfahrens darstellen, möchte ich in meinen Ausführungen näher eingehen.

Die verschiedenen Untersuchungen mit C-Tests im mutter- und fremdsprachlichen Bereich sowie im Zweitsprachenbereich mit ausländischen Schülern sind das gemeinsame Ergebnis einer zweijährigen interdisziplinären Zusammenarbeit. Der C-Test wurde zusammen von Ulrich Raatz und Christine Klein-Braley (Universität Duisburg) entwickelt und in einer Reihe von Untersuchungen erprobt (s. RAATZ & KLEIN-BRALEY, 1982). Auf einer Tagung der Interuniversitären Sprachtestgruppe (IUS) stieß ich 1982 in Duisburg zu dieser Gruppe. Inzwischen wurden C-Tests in deutscher, englischer, französischer, spanischer und hebräischer Sprache (z.T. von kooperierenden ausländischen Kollegen) konstruiert und in über 30 kleineren und größeren empirischen Untersuchungen mit gutem Erfolg erprobt. Zu verschiedenen Untersuchungsaspekten und -ergebnissen liegen eine Reihe von Veröffentlichungen sowie Vortragsmanuskripte vor, auf die an entsprechender Stelle verwiesen wird. Über den Stand der bisherigen gemeinsamen Forschungsbemühungen möchte

---

*Der vorliegende Beitrag wurde auf dem Symposium Sprachstandsmessung bei ausländischen Schülern, 10. bis 11. Februar 1984, in der Freien Universität Berlin referiert; er wird im Tagungsbericht veröffentlicht.

dieser Beitrag eine Übersicht geben. Weitere Testentwicklungen in finnischer, türkischer und griechischer Sprache sind angelaufen. Die Entwicklung des C-Tests ist aber keineswegs abgeschlossen. Deshalb soll am Ende der Darstellung auf noch offene Fragen und geplante Untersuchungen eingegangen werden.

C-Tests wurden zunächst zur Leistungsmessung in der Muttersprache oder in einer Fremdsprache konstruiert. Ich selbst habe versucht, dieses Verfahren auch für die Gruppe der ausländischen Schüler in der Bundesrepublik Deutschland nutzbar zu machen. C-Tests unterliegen dabei der Einschränkung, daß sie nicht für Schulanfänger geeignet sind, da die Testbearbeitung gewisse Voraussetzungen an die selbständige Lese- und Schreibleistung der Schüler stellt. Auch sind C-Tests keine diagnostischen Tests, die für didaktische Entscheidungen herangezogen werden können. Für eine Reihe von pädagogischen Entscheidungen zur Differenzierung ausländischer Schüler oder als bloße Information über den globalen Sprachstand liefern C-Tests aber mit ihrem Gesamtpunktwert eine wichtige Information für den Lehrer.

Bestrebungen zur Entwicklung von Sprachtests für ausländische Schüler haben in den letzten Jahren sehr zahlreich eingesetzt. Bisher liegen aber keine standardisierten Gruppentests vor. Vielmehr stellen die meisten Verfahren eine Art Lehrerhilfe für die individuelle Schülereinschätzung dar. Die linguistischen Grundlagen und die pädagogischen Zielsetzungen variieren dabei äußerst stark. Eine empirische Kontrolle der Verfahren wird meist vernachlässigt oder ganz vermieden, wobei unter anderem als Grund vorgegeben wird, man wolle keine diagnostischen Instrumente schaffen, mit denen den ausländischen Schülern und ihren Eltern die Erwägung der Rückkehr in ihre Heimat nahegelegt werden könnte (zur Kritik s. z.B. PREIBUSCH, HAGEMEISTER, SCHURICHT & SEYHAN, 1983 und SÜßMILCH & RAATZ, 1983). Mir erscheint umgekehrt eine viel größere Gefahr für die ausländischen Schüler von ungeprüften Testinstrumenten auszugehen, die - wie das Interesse von Schulbehörden und betroffenen Lehrern zeigt - oft recht unkritisch für die schulische Differenzierung und Förderung von Ausländerkindern herangezogen werden (sollen). Auf der Grundlage von Tests, die aber nur begrenzt objektiv durchführbar sind und den Sprachstand dieser Schüler kaum reliabel und valide erfassen, können natürlich auch keine verantwortungsbewußten pädagogischen Entscheidungen getroffen werden.

Die folgende Darstellung kommt in Punkt 5 auf den Einsatz von C-Tests bei ausländischen Schülern zurück. Sie will zeigen, daß - wie die bisherigen Untersuchungen andeuten - der C-Test eine linguistisch fundierte und psychometrisch abgesicherte Möglichkeit sein kann, zuverlässige Informationen über den allgemeinen Sprachstand ausländischer Schüler zu erheben, wodurch der C-Test zu einem interessanten Forschungsinstrument wird. In der Schule bietet er in Ergänzung zum Lehrerurteil die Basis für Treatmentzuweisungen im Zweitsprachenunterricht, was zu einer verbesserten Sprachausbildung beitragen kann. Neben diesem speziellen Anwendungsgebiet von C-Tests, sollen aber auch seine theoretischen Grundlagen, sowie Untersuchungsergebnisse mit anderen Stichproben, Validitätsuntersuchungen zum C-Test und die Vorhersage der Schwierigkeit von C-Tests umfassend behandelt werden.

2. Grundlagen der Sprachdiagnostik

2.1. Sprache als komplexes Kommunikationssystem

Will man valide Sprachtests konstruieren, so besteht die wichtigste Voraussetzung darin, das Phänomen, das erfaßt werden soll, zu verstehen. Sprachbeherrschung wird zunächst in vielen Situationen eher unbewußt getestet, etwa bei schulischen Aufgabenstellungen - selbst im Mathematikunterricht -, aber auch bei der Bearbeitung von Intelligenztests. Insbesondere in einer Fremdsprache oder einer Zweitsprache sind sprachliche Fähigkeiten eine grundlegende Bedingung für jegliche Form des schulischen Lernens und für soziale Kontakte.

LADO (1961), der sich als einer der ersten um die Verknüpfung linguistischer Erkenntnisse und psychometrischer Erfordernisse bemühte, beschreibt Sprache als ein komplexes Kommunikationssystem mit verschiedenen Schwierigkeitsebenen, die in komplizierter Weise miteinander verbunden sind. Jede Sprache besitzt dabei eine eigene Struktur, die ein in sich geschlossenes System darstellt. Jedem sprachlichen Vorgang liegt ein komplexer Sprachmechanismus zugrunde, der sich auf ein System von Gewohnheiten stützt, mit dem wir beispielsweise eine Auswahl an Wörtern vornehmen, die uns der Gesprächssituation oder dem Partner angemessen erscheinen. Wir ergänzen Flexionsendungen, strukturieren Sätze oder auch ein längeres Gespräch und verfolgen dabei eine bestimmte Gesprächsabsicht. Wir variieren möglicherweise die Betonung, modulieren den Klang der Stimme und unterstützen unsere Aus-sage durch Gestik und Mimik, wobei wir vielleicht unbewußt unsere Stimmung zu erkennen geben oder aber gezielt eine bestimmte Wirkung bei den Kommunikationspartnern anstreben, indem wir bereits Eindrücke der anderen antizipieren oder Reaktionen wahrnehmen und in unser sprachliches Verhalten einfließen lassen. LADO bezeichnet dementsprechend Sprache als komplexes Werkzeug des menschlichen Geistes, dessen Aufbau und Struktur es zu erkennen gilt, um Aufschlüsse über den Spracherwerb der Muttersprache oder die Erlernung einer Fremdsprache zu gewinnen. Die strukturalistische Linguistik stellte deshalb Listen von sprachlichen Einzelphänomenen zusammen, die die Grundlage für die Entwicklung moderner Sprachtests bildeten, die meist aus mehreren Untertests zu verschiedenen sprachlichen Fertigkeiten bestanden. Das Verständnis des Sprachlernprozesses war dabei stark durch den Behaviorismus geprägt, so daß die Beherrschung des Regelsystems einer Sprache in den Vordergrund rückte (pattern drill): nur wer die Regeln einer Sprache beherrscht, kann auch Sprache als Ganzes beherrschen.

Dieser Ansatz wurde in den sechziger Jahren insbesondere von CHOMSKY (1957, 1965) abgelehnt. Eine strukturalistische Beschreibung von Sprache konnte nach seiner Auffassung niemals vollständig sein; sie beachtete nicht die Kreativität von Sprache. Wäre Sprache auf bloße Reproduktion beschränkt, so könnte jemand, der eine Sprache lernt, keine neuen sprachlichen Äußerungen verstehen oder sie selbst formulieren. Nach CHOMSKY wird Sprache nach einem Regelsystem (der Grammatik) und dem Wortschatz (den Inhalten) ständig neu produziert. Mit Hilfe der 'Generativen Transformationsgrammatik' versuchte er Regeln zu erstellen, mit denen alle Sätze einer Sprache erzeugt werden können.

Eine Erweiterung erfuhr dieser Ansatz dadurch, daß die kommunikative Kompetenz als situationsbedingte sprachliche Handlungsfähigkeit eine besondere Bedeutung durch die (sozio)linguistische Forschung erhielt. So forderte HYMES (1964) die Ausweitung der linguistischen Fragestellung durch Einbeziehung unterschiedlicher Kontexte und die Hervorhebung der Rolle der Situation für das Sprachverhalten. Dadurch sollte der kommunikativen Kompetenz ein Platz neben der linguistischen Kompetenz eingeräumt werden. Ferner wurde - ausgehend von CHOMSKY - nun auch unterschieden zwischen sprachlicher Kompetenz, dem potentiellen Generieren von Sprachstrukturen, und sprachlicher Performanz, dem

aktuellen Produzieren von Texten (s. HUNDSNURSCHER, 1973). Der hier angesprochene Trend in der Linguistik - auch als kommunikative Wende bezeichnet - war natürlich dadurch bedingt, daß dem funktionalen Gebrauch von Sprache ein größerer Wert beigemessen wurde. SPOLSKY (1968) faßt diese Wende in die treffende Frage: "What does it mean to know a language, or how to get someone to perform his competence?"

2.2. Strukturalistische vs. integrative Sprachtests

Unterschiedliche linguistische Sprachmodelle erfordern auch unterschiedliche sprachdiagnostische Ansätze. Im strukturalistischen Sprachkonzept wird eine Unterscheidung zwischen Sprachebenen und Sprachfertigkeiten getroffen, so daß Sprache in verschiedene Elemente unterteilt wird. CARROLL (1968) weist in diesem Zusammenhang darauf hin, daß etwa der Zweitspracherwerb durch Wechselbeziehungen zur Muttersprache, durch langsameren Lernfortschritt oder durch mangelnde Motivation beeinflußt werden kann. Deshalb sollten die rezeptiven Fertigkeiten (Hören und Lesen) getrennt von den produktiven Fertigkeiten (Sprechen und Schreiben) getestet werden, da sie geringere Korrelationen als in der Muttersprache aufweisen. Strukturalistische Sprachtests sollten von unterschiedlichen sprachlichen Bereichen ausgehen, in denen individuelle Differenzen aufgespürt und gemessen werden können. Ähnlich wie CARROLL entwirft auch DAVIES (1977, S. 72) eine Sprachstrukturmatrix, die nach Sprachebenen (levels) und Sprachfertigkeiten (skills) untergliedert ist. Die Matrix dient als Hilfestellung für die systematische Konstruktion von Sprachtests; beispielhafte Erhebungstechniken werden in den Zellen der Matrix aufgeführt. Mit einer Vielzahl von Tests (bzw. Subtests) ist nun eine umfassende Erhebung der verschiedenen sprachlichen Aspekte möglich, so daß ein umfangreiches Profil der Sprachkompetenz eines Lernenden erstellt werden kann (vgl. hierzu etwa auch die Entwicklung des 'Allgemeinen Deutschen Sprachtests' (ADST) von STEINERT, 1978).

Gegen diese Zerstückelung von Sprache in strukturelle Einzelelemente wendet sich neben SPOLSKY (1968) insbesondere OLLER (1979), der auf die Diskrepanz hinweist, daß Sprache weder auf diese Weise gelernt noch gelehrt werden kann, womit strukturalistische Tests (discrete point tests) ihre diagnostische Funktion verlieren würden. GRIMM (1978, S. 364) stellt unter Hinweis auf die Entwicklung des Heidelberger Sprachentwicklungstests (H-S-E-T) ebenfalls heraus, "daß einzelne Sprachkomponenten nicht in 'reiner' Form gemessen werden können. Denn der Spracherwerb stellt keinen additiven, sondern vielmehr einen synergistischen Prozeß der Art dar, daß über den kommunikativen Gebrauch von Sprache das Kind erst ihren Inhalt und ihre Form lernt; die sprachliche Form (Syntax) wird über die Erfahrung mit Inhalten (Semantik) vermittelt und umgekehrt." Die Konsequenz dieser Kritik liegt in der Forderung nach ganzheitlichen Testverfahren.

SPOLSKY (1975, 1981) beschreibt die Entwicklungsgeschichte des modernen Fremdsprachentestens und stellt drei Trends heraus, die er als traditionelle (vorwissenschaftliche), moderne (psychometrisch-strukturalistische) und post-moderne (psycholinguistisch-soziolinguistische) Phasen bezeichnet. Die erste Phase ist eher gekennzeichnet durch Prüfungen als durch Tests, bei denen die Messung einzig auf dem Urteil des Prüfers basiert und auch kaum hinterfragt wird. Die Erhebungsverfahren waren Übersetzungen, Aufsatz und einige grammatische Übungen. Der Hauptkritikpunkt an diesen traditionellen Prüfungen war insbesondere deren mangelnde Reliabilität (s. hierzu z.B. SCHRÖTER, 1971; INGENKAMP, 1977).

In der zweiten, der psychometrisch-strukturalistischen Phase wurden deshalb objektive und standardisierte Tests entwickelt, um zuverlässige Messungen zu gewährleisten. Die Psychometrie entwickelte sich zu einer Spezialwissenschaft, die - so SPOLSKY - möglicherweise eine unheilige Allianz mit der strukturalistischen Linguistik einging. Es entstanden Tests zu einer Vielzahl sprachlicher Bereiche. Grundlage der Tests waren zumeist multiple-choice Items. "There was a great liking for measuring whatever could be easily quantified and easily measured, and a willingness to avoid the question of what it was one really was testing" (SPOLSKY, 1981, S. 16). Die Kritik spricht natürlich die mangelnde Validität der Tests der zweiten Phase an, die

oftmals einfach den reliablen Verfahren unterstellt wurde (zur Kritik s. u.a. OLLER, 1979; OLLER & PERKINS, 1980).

Die dritte, die post-moderne Phase in der Sprachtestentwicklung war das Ergebnis der Unzufriedenheit mit den vorhandenen Tests. Psycholinguisten stellten die Frage nach der Beziehung zwischen Kompetenz und Performanz, während Soziolinguisten die Variation im Sprachgebrauch hervorhoben. Damit wurde das Problem der Validität von Sprachtests in den Vordergrund geschoben. Tests sollten nicht länger die Kenntnis der sprachlichen Regeln überprüfen, sondern die Performanz, den funktionalen Gebrauch von Sprache messen. Sprache sollte nicht in Elemente unterteilt werden, vielmehr wurde versucht, Sprache mit ganzheitlichen, integrativen Testverfahren zu erfassen. Verfahren der ersten Phase wurden wieder aufgegriffen, doch wurden die Erfordernisse der Testgütekriterien aus der zweiten Phase weiterhin berücksichtigt. Diktat, Aufsatz und Interview gewannen wieder an Bedeutung. Aber auch ein neues Testverfahren, auf das noch ausführlicher eingegangen werden soll, wurde in den Vordergrund gestellt: der Cloze-Test, der eigentlich zur Bestimmung von Textschwierigkeiten erstellt worden war (s. TAYLOR, 1953; OLLER & CONRAD, 1971; OLLER, 1973).

2.3. Der Bezug zur diagnostischen Fragestellung

Discrete-point Tests sind das Ergebnis eines strukturalistischen Sprachkonzepts. Sie sind im engeren Sinne diagnostische Tests. "Die Informationen, die beim Diagnostiktest zu einzelnen Fertigkeiten gesammelt werden, dienen dem Lehrer zur Entscheidung über weitere Maßnahmen des Lehr- und Lernprozesses" (DAVIES, 1973, S. 32). Zumeist sollen Lernlücken ermittelt werden, die sich auf bestimmte Lerninhalte oder Unterrichtseinheiten beziehen. OLLER (1979) verweist in diesem Zusammenhang auf die enge Beziehung dieser Tests zu Sprachlerntheorien, insbesondere zu pattern drills, die aber im wesentlichen ohne Bedeutungs- und Kontextbezug sind und damit auch kaum Motivation für den kommunikativen Sprachgebrauch bedeuten. Unterstützt wurde diese Entwicklung durch die kontrastive Linguistik, die durch Sprachvergleich zwischen Erst- und Zweitsprache Kontrast- oder

Interferenzphänomene hervorhob und solche schwierigen Sprachmuster zum Hauptgegenstand für den Sprachunterricht und für Testitems machte.

Sprachdiagnostik auf der Grundlage strukturalistischer Tests steht in einem unmittelbaren Zusammenhang mit Sprachförderung. OLLER (1979) lehnt dagegen die Ansicht strikt ab, daß Sprachelemente isoliert gelehrt (oder getestet) werden können, ohne daß ein Kontextbezug hergestellt wird und ohne daß die Wechselwirkungen solcher Elemente in einem Kommunikationszusammenhang beachtet werden. Da wichtige sprachliche Eigenheiten innerhalb des strukturalistischen Sprachkonzepts verlorengingen, sei dieses Konzept uneffektiv als Grundlage für sprachlichen Unterricht und Sprachtesten. "The fact is that in any system where the parts interact to produce properties and qualities that do not exist in the parts separately, the whole is greater than the sum of its parts" (OLLER, 1979, S. 212). Eine gegensätzliche Auffassung vertritt SALIMBENE (1980). Da kommunikative Kompetenz im Kontext einer aktuellen Kommunikationssituation gemessen werden muß, bemängelt er: "Since each 'testee' might answer differently (and would if asked to perform the same communicative act at a different time), the entire concept of test validity and reliability is questionable" (zit. in STEVENSON, 1982, S. 15).

Als Resümee ist festzuhalten, daß das Lehren und Testen kommunikativer Kompetenz ein herausragendes Ziel der Sprachvermittlung und Sprachdiagnostik darstellt. Es erscheint fraglich, ob es das einzige Ziel ist, wie OLLER fordert. Zwar bedeutet die Beherrschung einzelner sprachlicher Elemente einer Sprache noch nicht, daß man sie auch funktional beherrscht, doch schließt dies nicht aus, daß der Sprachunterricht und das Sprachtesten auch andere, möglicherweise untergeordnete Ziele haben können. Es besteht die Gefahr, daß Sprachtests, nur weil sie vorgeben, bestimmte sprachliche Elemente zu messen, grundweg abgelehnt werden und ihr förderungsorientierter Ansatz damit wegfällt. Da die Kritiker strukturalistischer Tests aber, so argumentiert STEVENSON (1982), die Behauptung aufstellen, daß Sprache nicht in Elemente für Unterrichts- und Testzwecke untergliedert werden kann, so entkräftigen sie ihren eigenen Vorwurf gegen

discrete-point Tests, die dann natürlich auch keine isolierten sprachlichen Elemente messen. In Wirklichkeit besitzen ja auch Wortschatz- oder Grammatiktests durch ihre Instruktionen und die Itemvorgaben gewisse Beziehungen zur kommunikativen Kompetenz.

Die Gegensätze, die zwischen discrete-point Tests und integrativen/pragmatischen Tests bestehen, scheinen nicht dadurch gelöst werden zu können, daß man den einen oder den anderen Testansatz einfach ablehnt. Der Einsatz eines Tests ist vielmehr bedingt und gerechtfertigt durch die diagnostische Fragestellung. Strukturalistische Tests erhalten ihren besonderen Wert natürlich dadurch, daß sie Informationen für die Unterrichtsgestaltung und die individuelle Förderung liefern. Wird allerdings die vielfältige Testinformation nicht genutzt oder nicht benötigt, so stellt sich die Frage der Testökonomie. Tests in der Sprachausbildung sollten möglichst effektiv objektive diagnostische Informationen erheben, die pädagogisches Handeln rational begründet. Die Entscheidung für strukturalistische oder globale Testverfahren ist darüberhinaus eine Abwägung zwischen den Aspekten der Reliabilität und der Validität. Der Sprachdiagnostiker sollte im Hinblick auf die jeweilige Fragestellung das Für und Wider eines Testansatzes bedenken, insbesondere sollte die Bedeutung der Gütekriterien Reliabilität und Validität sowie die Nützlichkeit eines Tests diskutiert werden. STEVENSON (1982, S. 20f) unterstützt diese Forderung in seinem abschließenden Statement: "Progress in language testing is dependent upon a mutual effort among that society made up of language testers, test users, and test takers, and only when more careful attention is given to 'the subject matter being dealt with' can we expect more exactness in language testing".

3. Die Entwicklung globaler Sprachtests

3.1. Das Cloze-Prinzip

TAYLOR gilt als der Erfinder des Cloze-Prinzips. Er ist insbesondere verantwortlich für den mnemotechnischen Begriff 'cloze', der eine Abwandlung des Wortes 'close' darstellt. Aus einem Prosatext werden Wörter herausgenommen, so daß Lücken in einem Text zu 'schließen' sind. Lückentexte sind aber schon wesentlich früher vor allem zur Untersuchung von Lern- und Gedächtnisvorgängen herangezogen worden (s. EBBINGHAUS, 1897; Binet, 1905). TAYLOR selbst hat cloze-Tests zur Bestimmung von Textschwierigkeiten verwendet (1953; 1956). Aus Texten wurden entweder zufällig Wörter gestrichen oder es wurde nach einer festen Regel jedes n-te Wort getilgt (wobei n zwischen 5 und 10 lag). Die Lücken mußten nun von den Probanden wieder ausgefüllt werden, wobei nur exakte Lösungen akzeptiert wurden, die dem Originaltext entsprachen. TAYLOR stellte eine hohe Übereinstimmung mit Lesbarkeitsformeln fest. Er empfahl die systematische Streichung von jedem n-ten Wort und unterstellte, daß damit eine Zufallsstichprobe der sprachlichen Elemente getilgt würde.

CARROLL, CARTON & WILDS (1959) erprobten das Cloze-Prinzip auch in der Fremdsprache. SPOLSKY, SIGURD, SATO, WALKER & ARTERBURN (1968) entwickelten nach einem ähnlichen Prinzip einen Noise-Test, der aus einem Diktat mit Störgeräuschen bestand. Erst OLLER & CONRAD (1971) schafften aber den Durchbruch für das Cloze-Verfahren, wobei insbesondere für den Fremdsprachenbereich die Auswertungsmethode variiert wurde, so daß nun auch akzeptable Lösungen als Richtigantwort zugelassen wurden. Native speakers mußten dabei über die Akzeptabilität einer Lösung entscheiden.

Cloze-Tests messen - ähnlich wie Diktate - die Gesamtkompetenz in einer Erst-, Zweit- oder Fremdsprache. Der Grad, zu dem ein Proband richtige Antworten im Cloze-Test angibt, ist ein Index für die Leistungsfähigkeit einer Art von Sprachverarbeitungsmechanismus, der als 'general language proficiency' (SPOLSKY, 1968) oder 'pragmatic expectancy grammar' (OLLER, 1973) bezeichnet wird. Darunter ist eine internalisierte Grammatik zu verstehen, die den Sprachgebrauch steuert. Dieser Informationsverarbeitungsprozeß führt auf Seiten des Sprechers zu einer antizipatorischen Planung von Sprache, während auf Seiten des Zuhörers ständig Hypothesen gebildet werden, was wohl als nächstes geäußert wird. Auf diese 'Gewohnheiten' hatte schon LADO (1961) implizit hingewiesen. OLLER (1979) versteht darunter aber eher die aktive sprachliche Erwartungshaltung von Sprechern und Zuhörern. Der Ausprägungsgrad dieser Erwartungsgrammatik, der sich beispielsweise bei der Bearbeitung von Cloze-Tests zeigt, gilt als Indiz für die Gesamtkompetenz in einer Sprache.

3.2. Der C-Test - Eine Modifikation des Cloze-Prinzips

Aufgrund der einfachen Konstruktion und Auswertung und einer oftmals a priori unterstellten Gültigkeit hat der Cloze-Test eine weite Verbreitung erlangt. Zunehmend wurde aber auch Kritik am Cloze-Verfahren geäußert (ALDERSON, 1979; KLEIN-BRALEY, 1981). Die Kritikpunkte sind insbesondere:

1) Die getilgten Wörter eines Textes sind keine randomisierte Stichprobe der Wortarten. Dies führt u.a. dazu, daß Tests aus der gleichen Textvorlage bei unterschiedlichem Tilgungsanfang aber gleicher Tilgungsrate verschiedene Schwierigkeiten aufweisen und eine unterschiedliche Reliabilität und Validität besitzen.
2) Da Cloze-Tests wegen der Tilgungsrate und der Forderung nach einer angemessenen Anzahl von Items meist nur aus einem längeren Text bestehen, kann die Testfairneß beeinträchtigt sein. Man kann nicht erwarten, daß eine repräsentative Stichprobe der Sprache gesichert ist. Es fehlen aber Kriterien für die Auswahl geeigneter Texte.
3) Die Auswertungsmethoden werfen Probleme auf. Bei der exakten Auswertung sind Cloze-Tests selbst für kompetente Muttersprachler zu schwierig. Die akzeptable Auswertung ist dagegen sehr zeitaufwendig und weniger zuverlässig, da über die Akzeptabilität einer Lösung häufig Unklarheit besteht.
4) In Nachberechnungen ergab sich oft, daß Cloze-Tests weniger reliabel sind als allgemein unterstellt, insbesondere wenn die Probanden weniger heterogen waren. Die Paralleltestreliabilität erwies sich als gering.

Die Kritik, die insbesondere von KLEIN-BRALEY (1981) vorgetragen wurde, betrifft jedoch nur das Tilgungsprinzip und die Auswertungsmethode, nicht aber die theoretische Grundlage des Cloze-Verfahrens, die pragmatische Erwartungsgrammatik. Als mögliche Verbesserung des Cloze-Verfahrens wurde von JONZ (1976) und BROWN (1979) die Einführung von multiple-choice Items in Cloze-Tests vorgeschlagen.

RAATZ & KLEIN-BRALEY (1982; 1983) modifizierten das Tilgungsprinzip des Cloze-Tests. In einem authentischen Text wird nach einem einleitenden Satz jedes zweite Wort getilgt, jedoch nur noch die zweite Worthälfte. Bei Wörtern mit einer ungeraden Anzahl von Buchstaben wird ein Buchstabe mehr weggelassen. Wörter aus einem Buchstaben, wie sie im Englischen vorkommen, werden übersprungen. Mehrere solcher Texte mit 20 bis 25 Lücken werden zu einem C-Test zusammengestellt. Der Proband muß die fehlende Worthälfte ergänzen, wobei bisher nur fehlerfreie Ergänzungen akzeptiert wurden. Dieses neue Tilgungsprinzip, C-Prinzip genannt, gewährleistet, daß mehrere Texte mit mehr Items vorgegeben werden können. Die Auswertungsobjektivität ist gewährleistet, da es bis auf wenige Ausnahmefälle immer nur eine Lösung gibt. Im Gegensatz zu den Cloze-Tests sind C-Tests von kompetenten Muttersprachlern annähernd fehlerfrei zu bearbeiten. Außerdem ist gewährleistet, daß die getilgten Wörter in Bezug auf die Wortarten repräsentativ für den Ausgangstext sind. Die Abbildungen 1 und 2 zeigen jeweils einen Beispieltext aus einem englischen und einem deutschen C-Test.

Once upon a time, there was a little girl who lived with her mother, who was a widow. They we_____ so po_____ that o_____ day th_____ had not_____ left t_____ eat. T_____ little gi_____ went o_____ into t_____ woods t_____ play. S_____ was s_____ hungry th_____ she be_____ to c_____. An o_____ woman ca_____ up t_____ her. "W_____ are y_____ crying, m_____ child?" s_____ asked. "Bec_____ I am s_____ hungry," said the little girl. "Then you shall be hungry no more," said the old woman.

Abbildung 1: Beispielitem aus einem englischen C-Test (junior version)

Ein junger Hase, der zum erstenmal in die Sonne kam, sah plötzlich seinen Schatten neben sich. Hals üb_____ Kopf ran_____ der Ha_____ davon, do_____ das schw_____ Tier m_____ den lan_____ Hörnern ran_____ ebenso sch_____ neben i_____ her. D_____ Hase sch_____ Haken na_____ rechts u_____ links u_____ jagte i_____ den dun_____ Wald zur_____. Da end_____ war se_____ Verfolger verschwunden. Atemlos und erschöpft hielt der Hase an.

Abbildung 2: Beispielitem aus einem deutschen C-Test (für die Grundschule)

Die Texte stammen vorwiegend aus Lehrwerken (Lesebücher bzw. Sprachlehrwerke, Lektürehefte und Sachbücher), die der Lernstufe der Probanden entsprechen. Um eine gewisse Schwierigkeitsstreuung zu erzielen, werden aber auch Texte aus anderen Klassenstufen berücksichtigt. Die Testbearbeitungszeit liegt bei fünf bis sechs Texten bei etwa 30 Minuten. Die Auswertung ist wenig zeitaufwendig.

Für die empirische Analyse nach der klassischen Testtheorie werden die einzelnen Texte mit 20 bis 25 Lücken als sog. Superitems herangezogen. Eine Selektion oder Modifikation einzelner Lücken wurde bisher nicht durchgeführt. Die systematisch erzeugten Lücken eines Textes stellen, wie in Pilotstudien nachgewiesen wurde, eine repräsentative Sprachstichprobe für den Ausgangstext dar. Deshalb sollte dieses Prinzip beibehalten werden. Die Testanalyse kann allerdings dazu führen, daß sich einzelne Texte insgesamt als zu leicht, zu schwer oder zu wenig trennscharf erweisen. Dabei wird die Trennschärfe eines Textes durch die Korrelation des Teilrohwertes mit dem (part-whole-korrigierten) Gesamtrohwert berechnet. Neben dieser inneren Validität eines Textes läßt sich auch dessen äußere Validität parallel hierzu mit einem Außenkriterium bestimmen, z.B. mit Schulnoten. Die (instrumentelle) Reliabilität von C-Tests wird nach der Formel für CRONBACH's Alpha geschätzt. Zur Überprüfung der empirischen Validität wurden bisher teilweise verschiedene Noten oder globale Schätzurteile aus dem Sprachunterricht als Kriterium herangezogen. In einigen Untersuchungen konnten auch andere Sprachtests bzw. Sprachtestbatterien durchgeführt werden.

4. Der Einsatz von C-Tests im mutter- und fremdsprachlichen Unterricht

4.1. Die Erprobung von muttersprachlichen C-Tests

Zur Überprüfung der Praktikabilität von C-Tests, zur Kontrolle der ausgewählten Texte sowie zur Bestimmung der Reliabilität und Validität wurden eine Reihe von oftmals kleineren Untersuchungen zunächst mit deutschen und englischen C-Tests im muttersprachlichen Unterricht mit L1-Lernern durchgeführt (die Untersuchungsergebnisse sind z.T. ebenfalls dargestellt in RAATZ & KLEIN-BRALEY, 1983 und RAATZ, KLEIN-BRALEY & SÜßMILCH, 1983).

Englische C-Tests wurden bei zehn- bis fünfzehnjährigen Schülern in Yorkshire in insgesamt sieben Klassen erprobt. Dabei wurden zwei Testformen eingesetzt. Da es in England in der Primar- und Mittelschule keine Klassenwiederholung gibt, sind diese Stichproben sehr heterogen in ihrer Leistung, ganz im Gegensatz zu den älteren Schülern der Oberschule (ab etwa 12 Jahre). Tabelle 1 enthält die Untersuchungsergebnisse (Untersuchung Nr. 1 bis 6).

Untersuchung	Alter	N	r_{tt}	r_{tc}
1	10	44	.91	.73
2	11	15	.75	.86
3	12	30	.66	-
4	10	26	.88	.88
5	15	19	.77	.10
6	13	26	.92	.77

Tabelle 1: Reliabilitäts- und Validitätskoeffizienten von englischen C-Tests bei L1-Lernern

Die erstmals erprobten englischen Texte erreichen in allen Untersuchungen eine recht befriedigende Reliabilität. Die niedrige Reliabilität aus Untersuchung Nr. 3 ist dabei durch eine besonders homogene Klasse bedingt, da die Schüler aufgrund von Einstufungstests und Beurteilungen durch Lehrer zu einer Leistungsgruppe zusammengestellt wurden. Zur Bestimmung der Validität wurden Lehrereinschätzungen auf einer deutschen Notenskala erhoben, da es in England Schulnoten wie in Deutschland nicht gibt. Die Validitätskoeffizienten fallen mit einer Ausnahme überraschend hoch aus; sie sind dabei recht aufschlußreich. Für Untersuchung Nr. 3 kann kein Validitätskoeffizient angegeben werden, da die Lehrerin nach einem Monat Unterricht in dieser Klasse noch keine Schülereinschätzung vornehmen wollte. Für eine andere Lehrerin, die nach dieser kurzen Zeit dennoch die erbetenen Einschätzungen abgab (Untersuchung Nr. 5), fällt der Validitätskoeffizient äußerst niedrig aus. Die Lehrerin konnte die Englischleistungen ihrer Schüler demnach erst sehr ungenau beurteilen. In der sechsten Untersuchung haben die Schüler aus demselben Grunde ihre Leistung selbst eingeschätzt, was sie - gemessen am C-Test-Ergebnis - recht gut gemacht haben.

Die Untersuchungen mit deutschen C-Tests bei deutschen Schülern fanden im Bereich der Stadt Duisburg statt. Beteiligt waren 30 Klassen aus Grund- und Hauptschulen vom 3. bis zum 8. Schuljahr (Untersuchung Nr. 7 bis 12; siehe hierzu Tabelle 2).

Untersuchung	Klasse	N	r_{tt}	r_{tc}
7	3	130	.86	.80
8	4	92	.88	.72
9	6	63	.80	.47
10	7	49	.76	.47
11	8	62	.85	.60
12	3	110	.88	.52
		110	.87	.65

Tabelle 2: Reliabilitäts- und Validitätskoeffizienten von deutschen C-Tests bei L1-Lernern

Es wurden sechs Testversionen eingesetzt, wobei in der letzten Untersuchung zwei äquivalente Parallelformen erprobt wurden. Deshalb enthält die Tabelle für Untersuchung Nr. 12 jeweils zwei Werte. Validitätskriterium waren für die 6. bis 8. Klasse jeweils die Deutschnote, während für die Klassen 3 und 4 eine Deutschgesamtnote aus der Aufsatz- und Rechtschreibnote ermittelt wurde.

Die Reliabilitätskoeffizienten der untersuchten deutschen C-Tests sind insgesamt sehr zufriedenstellend, zumal die meisten Texte zum erstenmal erprobt wurden, und trotzdem kein einziger Teiltext ersetzt werden mußte. Die Übereinstimmungsvalidität der C-Tests mit Schulnoten ist durchweg hoch. Die untersuchten C-Tests messen also im wesentlichen das, was die Lehrer in ihren globalen Schulnoten beurteilen. Die Schwierigkeit sowohl der englischen als auch der deutschen C-Tests lag für alle Untersuchungen etwa im mittleren Bereich, wobei in den höheren Klassen die Schwierigkeit abnahm. Dies bedeutet eine langsame Annäherung an die native speaker Kompetenz an.

4.2. C-Tests in der fremdsprachlichen Ausbildung

Für Untersuchungen im fremdsprachlichen Bereich wurden englische C-Tests bei Duisburger Anglistikstudenten (LF-Lerner) eingesetzt. Die Ergebnisse zeigt Tabelle 3 (Untersuchung Nr. 13 bis 19).

Untersuchung	N	r_{tt}	r_{tc}
13	27	.81	.72
14	64	.80	.62
15	19	.91	.79
16	14	.77	.90
17	67	.91	.71
18	67	.90	.69
19	67	.84	.65

Tabelle 3: Reliabilitäts- und Validitätskoeffizienten von englischen C-Tests bei Anglistikstudenten (LF-Lerner)

Schwierigkeiten bereitete erstmals die Auswahl geeigneter Texte, da sich viele von ihnen als zu leicht erwiesen. Deshalb wurden insbesondere Sachtexte herangezogen. Die Reliabilitätskoeffizienten fallen bemerkenswert hoch aus. Als Validitätskriterium wurde jeweils der DELTA-Einstufungstest herangezogen (s. hierzu KLEIN-BRALEY & LÜCK, 1979). Dieser Test besteht aus einem Diktat sowie aus einer Reihe von Untertests mit Aufgaben zur Grammatik und zum Wortschatz. Die Korrelation zwischen den Ergebnissen des C-Tests und des Einstufungstests DELTA sind bemerkenswert hoch.

In zwei weiteren Untersuchungen (Nr. 20 und 21) wurden deutsche C-Tests mit ausländischen Jugendlichen und Erwachsenen durchgeführt, die an einem Sprachlehrinstitut (Eurozentrum in Köln) Deutschkurse für ein bis zwei Monate besuchten. Die Kursteilnehmer waren sehr heterogen und stammten aus den unterschiedlichsten Herkunftsländern. Ein eigener Einstufungstest des Eurozentrums lieferte das Validitätskriterium. Mit diesem Test, der etwa 70 Minuten dauert, werden Hörverständnis, Grammatikkenntnisse, sprachlicher Ausdruck und Leseverständnis erfaßt. In Untersuchung Nr. 20 wurde zunächst, um Erfahrung zu sammeln, die einfachste Testform für deutsche Schüler eingesetzt. Tabelle 4 zeigt die Untersuchungsergebnisse.

Untersuchung	N	r_{tt}	r_{tc}
20	99	.92	.82
21	153	.93	.85

Tabelle 4: Reliabilitäts- und Validitätskoeffizienten von deutschen C-Tests bei Studenten eines Sprachlehrinstituts (LF-Lerner)

Obwohl dieser Test eine mittlere Schwierigkeit aufweist und eine äußerst hohe Reliabilität und Validität erreicht, wurde der Inhalt der Texte als zu kindlich kritisiert. Deshalb wurden neue Texte von den Lehrern des Eurozentrums ausgewählt, die in Untersuchung Nr. 21 erstmals erprobt wurden. Auf Anhieb erreichte der so erstellte C-Test eine ähnlich hohe Reliabilität wie in Untersuchung Nr. 20. Auch die Übereinstimmung mit dem Eurotest ist so hoch, daß praktisch ein Paralleltest entwickelt wurde.

Um dem Vorwurf zu entgehen, die hohe Reliabilität sei bedingt durch die heterogene Stichprobe (wie für den Cloze-Test nachgewiesen), wurde der neu entwickelte Test zur Kontrolle einer homogenen Gruppe von Sprachschülern (englische Offiziere einer Armeesprachschule im höchsten Leistungskurs) vorgelegt. Doch auch bei dieser Stichprobe erzielte der Test eine Reliabilität von $r_{tt} = .93$.

Die Ergebnisse der C-Tests im LF-Bereich an Hochschulen und Sprachlehrinstituten sind äußerst befriedigend. Die Reliabilitäten sind sehr hoch, ebenso wie die Korrelationen zu den bisher verwendeten Einstufungstests. C-Tests könnten also ebenso als Einstufungstests verwendet werden und sind dabei wesentlich ökonomischer. Im Kölner Eurozentrum wird deshalb inzwischen ein C-Test innerhalb des Einstufungstests anstelle anderer Teiltests eingesetzt. Der Einsatz dieses C-Tests hat sich dort überaus bewährt und den Vorteil, daß zusätzliche Unterrichtszeit für die Kurse gewonnen wurde.

5. C-Tests im Zweitsprachenunterricht

5.1. Zur Situation der Sprachstandsmessung bei Ausländerkindern

Die schulische Ausbildung ausländischer Schüler in der Bundesrepublik Deutschland stellt eine besondere pädagogische Herausforderung dar. 843.500 ausländische Schüler besuchten im Schuljahr 1982/83 die allgemeinbildenden und berufsbildenden Schulen. Der bundesweite Anteil der ausländischen Schüler an der Gesamtzahl der Schüler stieg damit auf 7,7%. Dieser Anteil ist in den einzelnen Schultypen sehr unterschiedlich. Er beträgt in Grundschulen 14,1%, in Hauptschulen 13,1%, in Sonderschulen 11,2%, in Berufsschulen z.B. aber nur 4,6%. In Berlin, das die höchste Ausländerdichte hat, liegt dieser Anteil in den beiden erstgenannten Schultypen bereits bei etwa 40% bzw. darüber. Wie in

anderen Städten und Stadtteilen mit hoher Ausländerquote stellt sich hier natürlich die Frage nach einer besseren Differenzierung der ausländischen Schüler in Förderkurse oder verschiedene Klassenformen, um die schulische Ausbildung zu verbessern. Objektive Sprachtests könnten hierzu einen wichtigen Beitrag leisten.

In den vergangenen Jahren sind verschiedene Testentwicklungen oder erste Schritte hierzu meist aus der Praxis heraus begonnen worden. Diese Ansätze versuchen Sprachdiagnose und Sprachtherapie unmittelbar miteinander zu verknüpfen. Der Grad der Standardisierung der Verfahren und die empirische Kontrolle ihrer Anwendung ist aber oft ebenso unbefriedigend wie die Erfüllung linguistischer und testtheoretischer Erfordernisse. Einen umfassenderen Überblick über die Situation der Sprachtestentwicklung im Zweitsprachenbereich gibt SÜßMILCH (1984a; 1984b).

5.2. Der Einsatz von C-Tests bei ausländischen Schülern

Der C-Test ist wie andere globale Testverfahren kein diagnostischer Test. Er verlangt darüberhinaus Fertigkeiten im Lesen und Schreiben, was seine Einsatzmöglichkeit etwas einschränkt. Für eine Reihe pädagogischer Entscheidungen reicht aber bereits ein einziger Gesamtpunktwert für die allgemeine Sprachbeherrschung, etwa bei Klassifikationsentscheidungen wie sie gerade bei der inneren und äußeren Differenzierung von ausländischen Schülern anstehen. Für solche Fragestellungen sind strukturalistische Tests relativ unökonomisch, da die diagnostischen Informationen über jeden einzelnen Schüler möglicherweise gar nicht genutzt werden. C-Tests dagegen könnten sich für die Sprachstandsmessung bei ausländischen Schülern in der Zweitsprache Deutsch als sehr ökonomische Verfahren erweisen. Ihre Einsatzmöglichkeit wurde deshalb in verschiedenen Untersuchungen erprobt (s. hierzu SÜßMILCH & RAATZ, 1982; 1983).

Für eine erste Testerprobung mit ausländischen Grundschülern des dritten und vierten Schuljahres wurde auf bewährte Texte aus Untersuchungen mit deutschen Schülern zurückgegriffen. An den Untersuchungen Nr. 22 und 23 nahmen türkische und griechische Schüler aus Vorbereitungs- und Regelklassen teil. Schüler aus mehreren anderen Herkunftsländern bildeten eine dritte Gruppe. Daneben wurden aber auch die deutschen Schüler aus den Regelklassen mitgetestet. Elf von 31 sog. Seiteinsteigern (ausländische Schüler, die erst während des Schuljahres in die Bundesrepublik gekommen waren) konnten den Test nicht selbständig bearbeiten. Sie wurden von der Bearbeitung ausgeschlossen.

Die ausländische und auch die deutsche Stichprobe erwiesen sich als sehr heterogen, was sich darin zeigte, daß jeweils nahezu die gesamte Skala (0 bis 100 Punkte) ausgenutzt wurde. Der eingesetzte C-Test bzw. einige Textteile waren vor allem in der dritten Klasse zu schwer. Wie die Tabelle 5 aber zeigt, erreichte der C-Test in allen Stichproben eine sehr hohe Reliabilität.

Untersuchung	Nationalität	Klasse	N	r_{tt}	r_{tc}
	ausländisch	3	231	.86	.72
	türkisch	3	92	.85	.61
22	griechisch	3	107	.85	.74
	andere	3	32	.83	-
	deutsch	3	109	.92	.80
	ausländisch	4	245	.92	.69
	türkisch	4	94	.87	.41
23	griechisch	4	96	.92	.82
	andere	4	55	.88	.73
	deutsch	4	88	.88	.82

Tabelle 5: Reliabilitäts- und Validitätskoeffizienten eines deutschen C-Tests bei ausländischen und deutschen Schülern (L2- bzw. L1-Lerner)

Als Validitätskoeffizient jeder Stichprobe ist in Tabelle 5 jeweils der Median der pro Klasse bestimmten Validität angegeben. Damit sollte der klasseninterne Maßstab der Lehrerurteile ausgeglichen werden. Da Schülergruppen unter acht Schülern nicht berücksichtigt wurden, kann

ein Validitätswert nicht berechnet werden. Außenkriterien für die Bestimmung der Validität waren Einschätzungen der schriftlichen und mündlichen Leistung im Fach Deutsch auf einer vorgegebenen Skala (Skalierung von -5 bis +5), die in Tabelle 5 zu einem Kriterium zusammengefaßt sind. Die niedrige Validität für die türkischen Schüler ist mitbedingt durch die geringe Varianz ihrer Testwerte. In der vierten Klasse macht sich zusätzlich für die türkischen Schüler bemerkbar (ähnlich wie in Untersuchung Nr. 5), daß einige Vorbereitungsklassen erst vor acht Wochen von den jetzigen Lehrern übernommen wurden, und die Lehrer den Sprachstand dieser Schüler noch nicht zuverlässig beurteilen können. Ansonsten stimmen die C-Test-Ergebnisse recht hoch mit den Schätzurteilen der Lehrer überein.

Die Ergebnisse waren natürlich sehr ermutigend. Der in den Untersuchungen Nr. 22 und 23 eingesetzte C-Test, dessen Texte sich ja bei deutschen Schülern schon bewährt hatten, erfaßte den Sprachstand der ausländischen Schüler für alle Nationalitäten sehr genau. Die geringere Reliabilität in der 3. Klasse ist darin begründet, daß der untersuchte C-Test hier zu schwierig ist, während er für die deutschen Schüler dieser Klassenstufe angemessen erscheint. In der 4. Klasse ändert sich das Bild etwas. Jetzt ist derselbe C-Test für die deutschen Schüler eher zu leicht, während er für die ausländischen Schüler, mit Ausnahme der türkischen Stichprobe, genau die richtige Schwierigkeit besitzt. Es zeichnet sich demnach ab, daß je nach Lernfortschritt in der Zweitsprache Deutsch schülerspezifische Texte zu einem C-Test zusammengestellt werden müssen. Insbesondere für Schüler mit niedrigem Sprachstand sollte darüberhinaus versucht werden, leichtere Texte, eventuell unter Abwandlung des C-Prinzips, zu entwickeln. Auch scheint eine Testverkürzung möglich. Wie SÜßMILCH & RAATZ (1982) zeigen konnten, bringt selbst eine extreme Testverkürzung von fünf auf zwei Texte nur eine geringe Reliabilitätsminderung mit sich.

Für weitere Untersuchungen mit ausländischen Schülern wurden deshalb neue C-Tests entwickelt. In den Untersuchungen Nr. 24 und 25 sollten aber zunächst einmal neue Texte erprobt werden. Das C-Prinzip wurde deshalb beibehalten mit der Ausnahme, daß die Anzahl der fehlenden Buchstaben durch Striche angedeutet wurde. Voruntersuchungen mit englischen C-Tests bei Anglistikstudenten hatten bestätigt, daß Texte dadurch insgesamt leichter wurden, denn es bleibt meist nur noch eine einzige Lösungsmöglichkeit übrig, die schneller gefunden werden kann.

Dieser neuentwickelte C-Test wurde griechischen Schüler aus dritten und vierten Schuljahren vorgelegt. Diese Schüler hatten auch an der obigen Untersuchung teilgenommen. Die Texte besaßen in der 3. Klasse eine mittlere Schwierigkeit, in der 4. Klasse waren sie sogar etwas zu leicht. An den Untersuchungen Nr. 24 und 25 haben jeweils Schüler aus 3 Klassen teilgenommen (s. Tabelle 6).

Untersuchung	Nationalität	Klasse	N	r_{tt}	r_{tc}
24	griechisch	3	52	.82	.58 - .74
25	griechisch	4	50	.85	.66 - .90

Tabelle 6: Reliabilitäts- und Validitätskoeffizienten eines deutschen C-Tests bei griechischen Schülern (L2-Lerner)

Die Ergebnisse bestätigen, daß dieser Test auf Anhieb eine recht hohe Reliabilität erzielt, die bei Ausschluß wenig trennscharfer Texte sogar noch ansteigt. Validitätskriterium war die Beurteilung der Deutschleistung aus den Untersuchungen Nr. 22 und 23. In der Tabelle ist der niedrigste und höchste Wert der je Klasse ermittelten Validität je Schulstufe angegeben. Selbst mit dem vier Monate zurückliegenden Lehrerurteil besteht eine hohe Übereinstimmung. Da erstmals Schüler zwei C-Tests innerhalb weniger Monate bearbeitet haben, konnte die Übereinstimmung der Testergebnisse überprüft werden. Die Korrelation der beiden C-Test-Ergebnisse beträgt für die 3. Klasse r = .81 und in der 4. Klasse r = .90. Diese hohe Übereinstimmung überrascht, zumal der zweite Test neue Texte mit einem etwas abgewandelten C-Prinzip enthielt und vier Monate Zwischenraum zwischen den Erhebungen lag.

In beiden Fällen wurden keine Texte überarbeitet oder ausgewechselt. Trotzdem wurden praktisch ad hoc Paralleltests konstruiert.

Um den globalen Sprachstand von Schülern bereits am Ende des zweiten Schuljahres erfassen zu können, wurde ein weiterer C-Test konstruiert. Hierzu wurden kürzere Texte aus Erstlesewerken und aus Deutschlehrwerken für Ausländerkinder zu einem Test zusammengestellt (sechs Texte mit nur noch jeweils zehn Lücken). Auch das C-Prinzip wurde weiter abgewandelt, denn es wurde, um die Schwierigkeit der Texte noch weiter zu senken, nicht nur die Anzahl der fehlenden Buchstaben angedeutet, sondern auch nur noch jedes dritte Wort zur Hälfte weggelassen (s. hierzu BORN, 1984). Der Test wurde türkischen Schülern aus drei Vorbereitungsklassen gegen Ende des zweiten Schuljahres vorgelegt. Die Bearbeitungszeit wurde in dieser Untersuchung nicht beschränkt; sie betrug im Schnitt 40 Minuten, schwankte aber stark und lag zwischen 13 und 60 Minuten. Nur ein Schüler konnte den Test nicht bearbeiten, er war erst seit sechs Wochen in der Bundesrepublik. Die Ergebnisse dieser Untersuchung (Nr. 26) sind in Tabelle 7 wiedergegeben.

Untersuchung	Nationalität	Klasse	N	r_{tt}	r_{tc}
26	türkisch	2	41	.94	.72 - .88
27	türkisch	3	60	.84	.63 - .87

Tabelle 7: Reliabilitäts- und Validitätskoeffizienten einer Retestuntersuchung mit einem deutschen C-Test bei türkischen Schülern (L2-Lerner)

Der C-Test besitzt eine äußerst hohe Reliabilität. Sie sinkt selbst bei Verkürzung des C-Tests auf drei Texte mit nur noch 30 Lücken auf lediglich r_{tt} = .91 ab. Das Validitätskriterium wurde auf gleiche Weise erhoben wie in den Untersuchungen Nr. 22 und 23. In der Tabelle ist angegeben, in welchem Bereich die Validitätskoeffizienten der einzelnen Klassen liegen. Die Übereinstimmung mit dem Lehrerurteil ist in allen Fällen sehr hoch, sie ist in der Tendenz in leistungsschwächeren Klassen niedriger.

Eine Testwiederholung (Untersuchung Nr. 27; s. Tabelle 7) Mitte des 3. Schuljahres (mit 34 Schülern der ersten Erhebung) mit demselben, aber auf vier Texte verkürzten Test zeigt ähnliche Ergebnisse. Die Übereinstimmung beider Testergebnisse liegt bei $r_{tt'} = .85$ (Retest-Reliabilität). Die prognostische Validität des am Ende des 2. Schuljahres durchgeführten C-Tests mit der nach einem halben Schuljahr erhobenen Deutschnote beträgt $r_{tc'} = .71$ (bei allerdings nur $N' = 22$).

Abschließend kann herausgestellt werden, daß die ersten Erprobungen von C-Tests bei L2-Lernern sehr zufriedenstellende Ergebnisse erbracht haben. C-Tests können - zumindest in ihrer schriftlichen Form - frühestens ab Mitte des zweiten Schuljahres eingesetzt werden. Sie haben sich aber als praktikabel erwiesen und sind sehr ökonomisch, sowohl für die Testbearbeiter als auch für den Auswerter. C-Tests sind hochreliabel und stimmen hoch mit dem Lehrerurteil überein. Dies verdient umso mehr Beachtung, als das Lehrerurteil sonst oft als unzuverlässig herausgestellt wird, was den Validitätskoeffizienten eher mindert. In den verschiedenen Untersuchungen zeigte sich aber lediglich, daß nur in Klassen, die ein Lehrer erst vor kürzerem übernommen hat, der Validitätskoeffizient deutlich absinkt. Dies deutet daraufhin, daß C-Tests eine wichtige Hilfe für die frühe Einschätzung der sprachlichen Leistung etwa bei der Zusammenstellung von Leistungsgruppen bedeuten können.

6. Zur faktoriellen Validität von C-Tests

6.1. Die Beziehung zwischen Sprachstruktur und Sprachtests

Die Beziehung von Sprachtheorie, Sprachstruktur und Sprachleistungsmessung wurde bereits im zweiten Kapitel angesprochen. Der strukturalistische und der integrative Ansatz stehen dabei auch heute noch konkurrierend nebeneinander. "Jedes (der beiden Sprachmodelle) ist theoretisch fundiert, jedes ist empirisch belegt, jedes hat zur Entwicklung von speziellen Tests geführt, die im Rahmen des jeweiligen Modells reliabel und valide messen" (RAATZ, 1981, S. 3).

Um die Angemessenheit eines Modells zu kontrollieren, hat beispielsweise OLLER (1979) Hypothesen zur Sprachstruktur aufgestellt und sie dann empirisch geprüft. Ziel der faktorenanalytischen Studien ist es, Interkorrelationen zwischen verschiedenen Sprachtestergebnissen zu analysieren, indem eine ökonomische und gut interpretierbare Strukturierung der Daten gesucht wird. Vielfach haben diese Untersuchungen Hinweise auf einen allgemeinen Sprachfaktor (g-Faktor) ergeben (s. z.B. SANG & VOLLMER, 1978). Es traten aber in anderen Untersuchungen auch zusätzliche Gruppenfaktoren auf oder der g-Faktor ließ sich nicht isolieren. Den jeweiligen Ergebnissen entsprechend wurden dann jeweils Forderungen nach strukturalistischen oder globalen Tests gestellt.

Die unterschiedlichen Ergebnisse sind natürlich bedingt dadurch, daß jeweils unterschiedliche Tests und unterschiedlich homogene Stichproben berücksichtigt werden. Ferner ist das methodische Vorgehen bei der Faktorenanalyse sehr uneinheitlich. Die Faktorenanalyse ist ein heuristisches, hypothesengenerierendes Verfahren; es ist eine deskriptive Methode zur Überprüfung der Dimensionalität komplexer Merkmale (s. BORTZ, 1977). Es sollte zur Bildung nicht aber zur Testung von Hypothesen eingesetzt werden, weil man sonst einem Zirkelschluß unterliegt. "Sprachtheorien sollte man auf andere Weise überprüfen" (RAATZ, 1981, S. 24).

6.2. Untersuchungen zur konvergenten, divergenten und faktoriellen Validität von C-Tests

Die konvergente Validität von C-Tests ist oben bereits in den verschiedenen Untersuchungen bestätigt worden. C-Tests korrelieren hoch mit Schulnoten des entsprechenden Faches, mit Lehrereinschätzungen zu sprachlichen Leistungen sowie mit anderen Sprachtests.

Zur Beschreibung der Struktur von C-Tests berichten RAATZ & KLEIN-BRALEY (1983) über Faktorenanlyse zu C-Tests. In der oben beschriebenen Untersuchung Nr. 18 mit Anglistikstudenten (LF-Lernern) wurde der Hochschuleinstufungstest DELTA (mit sechs Subtests) durchgeführt, sowie zusätzlich der English Progress Test F3, der die Beherrschung der Muttersprache bei 12- bis 13jährigen englischen Schülern mißt. Es ergab sich ein erster Faktor, der 87% der Varianz aufklärte. Die höchste Ladung besaß der F3 mit .91, die niedrigste der DELTA-Subtest 'Zeitformen von Verben' mit .67. Der C-Test lag bei .72.

In den Untersuchungen Nr. 21 und 22 bei erwachsenen, ausländischen Deutschlernern wurde ebenfalls zur Validitätsprüfung eine Testbatterie eingesetzt (Eurotest mit 4 Untertests). Faktorenanalysen nach der Hauptachsenmethode führten in beiden Untersuchungen zu nur einem allgemeinen Faktor, der 71% bzw. 70% der Gesamtvarianz aufklärte. Mit diesem gemeinsamen globalen Sprachfaktor lassen sich sowohl die verschiedenen Untertests als auch der C-Test beschreiben. Die beiden sehr unterschiedlichen C-Tests liefern fast identische Ergebnisse. Hohe Ladungen zeigen die Untertests Hörverstehen, Grammatik und Ausdruck, doch hat der C-Test mit .88 und .90 jeweils die höchsten Ladungen. Eine deutlich geringere Ladung hat dagegen jeweils der Untertest Leseverständnis.

Über eine umfassende Untersuchung zur konvergenten, divergenten und faktoriellen Validität von C-Tests berichtet RAATZ (1984). Die mögliche Kritik an den eben beschriebenen Faktorenanalysen, wo teilweise sehr heterogene Stichproben und ähnliche Tests schon eine gute Voraussetzung für einen g-Faktor bedeuten, läßt sich so überprüfen. In einer Studie mit deutschen C-Tests (in zwei Parallelformen) wurden 75 deutsche Schüler des 5. Schuljahres untersucht (jeweils eine Klasse einer Hauptschule, einer Realschule und eines Gymnasiums). Folgende Variablen wurden in die Untersuchung einbezogen: C-Test (Version A und B), Schulnoten in Deutsch, Englisch und Mathematik, eine Lehrereinschätzung zu den Grammatikkenntnissen, der Diagnostische Test Deutsch (DTD 4-6), der Rechtschreibtest DRT 4-5, der sprachfreie Intelligenztest Frankfurter Denkaufgaben (FDA 3-6) und ein Konzentrationstest (test d2). Der DTD 4-6 besitzt sechs Untertests: Passiver Wortschatz (PW), Analogiefindung (AF), Textstrukturierung (TS), Instruktionsverständnis (IV), Leseverständnis (LV), Aktiver

Wortschatz (AW).

Zur Prüfung der konvergenten und divergenten Validität der beiden C-Tests, die jeweils von der Hälfte der Stichprobe bearbeitet worden waren, wurden die Korrelationen zu allen anderen Variablen bestimmt. Dabei wurde der klasseninterne Maßstab der Note durch eine Korrektur und die Ergebnisse der C-Tests durch Transformation in T-Werte vergleichbar gemacht. Daher sind die beiden Substichproben in Tabelle 8 sowohl getrennt als auch zusammengefaßt aufgeführt.

C-Test (A und B)		C-Test (A)		C-Test (B)	
DRT	.68	DRT	.63	TS	.76
TS	.66	PW	.62	DRT	.75
Deutsch	.61	TS	.57	Deutsch	.65
Grammatik	.60	Deutsch	.57	Grammatik	.64
Mathematik	.57	Grammatik	.57	Mathematik	.64
Englisch	.54	Englisch	.57	FDA	.57
IV	.54	IV	.54	Englisch	.54
FDA	.51	Mathematik	.49	IV	.53
PW	.49	FDA	.46	LV	.50
AF	.44	AF	.42	AF	.46
LV	.43	AW	.39	AW	.40
AW	.40	LV	.36	PW	.36
d2	.20	d2	.25	d2	.14

Rangkorrelation: r_{AB} = .63 / r'_{AB} (ohne PW) = .87

Tabelle 8: Korrelationen zwischen C-Tests und anderen Variablen für eine deutsche Stichprobe (L1) (aus: RAATZ, 1984)

Der C-Test besitzt die höchsten Korrelationen zum Rechtschreibtest sowie zum Untertest Textstrukturierung, hohe Korrelationen zu Deutsch und Grammatik, mittlere Korrelationen zu Mathematik und allgemeiner Intelligenz, geringere Korrelationen zu Leseverständnis und Wortschatz und nur eine minimale Korrelation zum Konzentrationstest. Zwischen den beiden Substichproben ist eine hohe Übereinstimmung festzustellen. Einzig der Untertest Passiver Wortschatz zeigte sehr unterschiedliche Beziehungen zu den beiden C-Test-Versionen. Ohne diesen Subtest besteht eine Rangkorrelation von $r'_{AB} = .87$. Die Daten enthalten folgende Tendenz: Der C-Test mißt hauptsächlich die allgemeine Sprachbeherrschung, zu einem geringeren Maß aber auch Aspekte der Intelligenz. Das C-Test-Ergebnis ist relativ unabhängig von der Konzentrationsfähigkeit. Die anschließende Faktorenanalyse ergab zwei Faktoren, die 61% der Varianz der Gesamtstichprobe aufklärten (zum Vorgehen s. RAATZ, 1984). Tabelle 9 zeigt die Faktorenmatrix, wobei Ladungen größer .60 hervorgehoben sind.

Variable	Faktor I	Faktor II
DRT	.87 o	-.04
TS	.80 o	.22
Grammatik	.77 o	.32
Englisch	.74 o	.25
Deutsch	.73 o	.27
C-Test	.71 o	.33
d2	-.11	.74 o
Mathematik	.58	.65 o
FDA	.48	.63 o
IV	.43	.63 o
AF	.39	.61 o
LV	.49	.20
PW	.46	.26
AW	.23	.39

Tabelle 9: Faktorenmatrix für die Gesamtstichprobe (aus: RAATZ, 1984)

Der erste Faktor kann interpretiert werden als allgemeine Sprachbeherrschung. Er besitzt hohe Ladungen für den DRT, für Textstrukturierung, für die Noten aus dem Sprachunterricht und für den C-Test. Der zweite Faktor zeigt hohe Ladungen für den Konzentrationstest, für sprachfreie Intelligenz und für die Mathematiknote. RAATZ bezeichnet

ihn als Faktor des logischen Denkens. Auf diesem Faktor hat der C-Test nur eine niedrige Ladung. Die Faktorenstruktur ließ sich annähernd auch für die beiden Substichproben in einer Kreuzvalidierung nachweisen. Wenn die Ergebnisse auch in der Tendenz zeigen, daß C-Tests für L1-Lerner hauptsächlich die allgemeine (muttersprachliche) Sprachbeherrschung messen, so hat die Untersuchung insbesondere auch wegen der Stichprobengröße doch eher explorativen Charakter (vgl. die obigen Kritikpunkte).

7. Vorhersage der Schwierigkeit von C-Tests

Die Auswahl geeigneter Texte für C-Tests stellt ähnlich wie für Cloze-Tests ein Problem dar. Sowohl in Bezug auf den Inhalt als auch auf die Schwierigkeit eines Textes gibt es bisher kaum nützliche Kriterien. KLEIN-BRALEY (1984) fordert für diese Texte, daß sie keine speziellen Vorkenntnisse des Probanden enthalten sollen. Diese mögliche Testunfairneß wird im C-Test durch mehrere verschiedene Textvorgaben bereits ausgeglichen. Die einzelnen Texte sollten möglichst keine literarischen Texte sein, sondern einfache neutrale Erzähltexte ohne Fachterminologie.

Wie die Untersuchungen mit ausländischen Schülern gezeigt haben, besteht aber ein Bedürfnis nach schülerspezifischen (einfacheren oder schwereren) Texten. Gerade am Anfang des Sprachlernprozesses, aber auch auf einer höheren Stufe der Sprachbeherrschung (wie die Untersuchungen mit Anglistikstudenten gezeigt haben) müssen Texte gesucht werden, die ein entsprechendes Anforderungsniveau (eine spezifische Textschwierigkeit) besitzen. RAATZ & KLEIN-BRALEY (1983) haben für Beispieltexte nachgewiesen, daß die Schwierigkeit ein und desselben Textes, wenn man ihn L1- oder L2-Lernern vorlegt, mit dem Alter bzw. der Klassenstufe abnimmt. Diese Beziehungen scheinen linear zu verlaufen. Hat man also Vorinformationen über die Textschwierigkeit, so kann man mittels einer Regressionsgeraden die Schwierigkeit auch in anderen Klassen bestimmen. Der Verlauf der Regressionsgeraden für die Schwierigkeit von fünf Beispieltexten aus dem L1-Bereich in Abhängigkeit von der Klassenstufe (s. Abbildung 3) läßt darüberhinaus erkennen, daß die maximale muttersprachliche Kompetenz von Hauptschülern (bzgl. der untersuchten Texte des C-Tests) zwischen dem 9. und 12. Schuljahr erreicht wird. Bereits CARROLL (1961) hatte aufgrund seiner Untersuchungen vermutet, daß der Lernprozeß in der Muttersprache (Alltagssprache) frühestens in der Pubertät abgeschlossen ist.

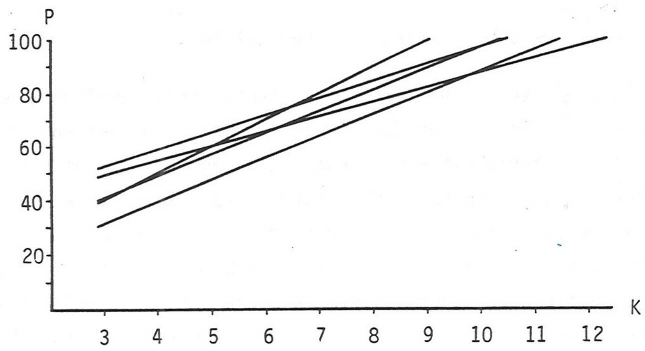


Abbildung 3: Die Textschwierigkeit in Abhängigkeit von der Klassenstufe bei L1-Lernern (aus: RAATZ & KLEIN-BRALEY, 1983, S. 134)

Für die Erstellung von C-Tests ist es aber sinnvoll, bereits vorab Informationen über die Textschwierigkeit zu besitzen, da die Texte eines C-Tests in etwa eine mittlere Schwierigkeit besitzen sollten, auf der anderen Seite aber auch eine gewisse Schwankung in ihrer Schwierigkeit aufweisen müssen. Bisher haben wir uns meistens darauf verlassen können, daß die Autoren von Lehrwerken uns die geforderten Schwierigkeiten liefern. Um aber objektive Kriterien für Textschwierigkeiten zu erhalten, beschäftigt sich KLEIN-BRALEY (1983) mit den Möglichkeiten und Grenzen, die aus der bisherigen Verständlichkeitsforschung (Readability Research) abgeleitet werden können. Insbesondere Lesbarkeitsformeln deuten die Möglichkeit an, daß der Versuch lohnend

sein könnte, sprachstatistische Indizes und die empirisch bestimmte Schwierigkeit von C-Tests aufeinander zu beziehen. Da nahezu alle Formeln einen Index für syntaktische Komplexität (z.B. Satzlänge) sowie einen Index für Wortschatzvielfalt (oft bezogen auf Worthäufigkeitslisten, was problematisch ist) enthalten, übernahm auch KLEIN-BRALEY diese Indizes in ihre Untersuchungen.

In Vorstudien erwiesen sich die Anzahl der Sätze im Text und die Type-Token-Ratio (TTR) als die beiden besten Prädiktoren für die Textschwierigkeit. Die TTR stellt dabei ein sprachunabhängiges Maß für die Wortschatzvielfalt dar. Sie wird berechnet, indem die Anzahl der Types (verschiedene Wörter) durch die Anzahl der Tokens (Summe der Wörter) dividiert wird. Die beiden Prädiktoren wurden in die Regressionsgleichung zur Bestimmung der Textschwierigkeit übernommen. Später wurde der Prädiktor Anzahl der Sätze durch die durchschnittliche Anzahl der Wörter je Satz ersetzt, denn die Bestimmung der Länge eines vorzugebenden Textes war oft inhaltlich und damit auch subjektiv beeinflußt.

KLEIN-BRALEY (1984) nennt mehrere Regressionsgleichungen für die Vorhersage der Schwierigkeit von deutschen Texten für verschiedene muttersprachige Stichproben. Die Gleichungen differieren wegen der unterschiedlichen Zielgruppen nur in einer additiven Konstante. Die Prädiktoren sind mit identischen Beta-Gewichten in den Regressionsgleichungen enthalten. In den Untersuchungen zeigen die vorhergesagte Textschwierigkeit und die empirisch ermittelte Schwierigkeit eine äußerst hohe Übereinstimmung (Rangkorrelation). Eine Ausnahme bilden die Klassen 9 bis 11 am Gymnasium; hier besitzen die Texte eine äußerst geringe empirische Schwierigkeit. Insgesamt ist es aber möglich, zufriedenstellende Schwierigkeitsvorhersagen für L1-Lerner aus dritten bis zehnten Klassen zu machen. Mit derselben Regressionsgleichung konnte ebenso die relative Schwierigkeit der Texte für L2- und LF-Lerner bestimmt werden. Auch die Vorhersage der Schwierigkeit englischer Texte ist über eine ähnliche Regressionsgleichung möglich. Lediglich der Prädiktor durchschnittliche Anzahl von Wörtern je Satz wurde dabei durch den Prädiktor durchschnittliche Satzlänge in Silben ersetzt.

8. Diskussion der Ergebnisse und Ausblick

Im zweiten Kapitel wurde die enge Beziehung zwischen Sprachtheorie, Sprachstruktur und Sprachleistungsmessung angesprochen. Sowohl im Bereich der Sprachtheorie als auch im Bereich der Sprachtests stehen heute konkurrierende Ansätze nebeneinander. Selbst empirische Untersuchungen bringen keine entscheidende Klarheit, wie die angesprochenen faktorenanalytischen Studien in Abschnitt 6.1 gezeigt haben. SPOLSKY (1981) hat die wechselnden Trends der Sprachdiagnostik beschrieben, die einmal die Struktur der Sprache, einmal deren integrative Komplexität betonen. Mir scheint die Entscheidung zwischen diesen Theorien und den damit verbundenen Testansätzen letztlich nicht lösbar. Einzelne Sprachkomponenten können nicht in reiner Form erfaßt werden; andererseits liefert die strukturalistische Linguistik wichtige Hinweise für den eigentlichen diagnostischen Aspekt von Sprachtests und für die Sprachvermittlung. Beide Ansätze scheinen ihre Rechtfertigung zu haben und sich in gewisser Weise zu ergänzen. Vielleicht sind deshalb die empirischen Untersuchungsergebnisse zur Angemessenheit der einen oder der anderen Sprachtheorie auch so diffus.

Sprache ist ein komplexes Phänomen, über dessen Struktur - ähnlich wie bei der Intelligenz - keine Einigkeit besteht. Vor diesem Hintergrund fällt dem Diagnostiker die Aufgabe zu, insbesondere den Anwendungsaspekt von Sprachtests zu berücksichtigen. Insofern muß je nach Fragestellung der eine oder andere Aspekt von Sprache stärker hervorgehoben werden, wobei die psychometrischen Erfordernisse an Sprachtests zu beachten sind. Wie die Entwicklungsgeschichte der Sprachtests gezeigt hat, wurde die Befürwortung strukturalistischer oder integrativer Testverfahren auch dadurch hervorgerufen, daß einseitig Reliabilität bzw. Validität der Verfahren angestrebt wurde, wobei das jeweils andere Kriterium vernachlässigt wurde. Wichtig ist aber, daß beide Kriterien Beachtung finden. "Es liegt eine gewisse Kunst darin, einen Test sowohl möglichst reliabel wie auch zugleich möglichst valide zu gestalten" (LIENERT, 1969, S. 294f).

Ein weiterer wichtiger Aspekt der Sprachdiagnostik betrifft die Frage der Ökonomie und der Nützlichkeit eines Testverfahrens. Wird diagnostische Information für den Lehr-Lern-Prozeß nicht benötigt bzw. nicht berücksichtigt, so können globale Testverfahren verwendet werden, deren Gesamtpunktwert eine ausreichende Grundlage für pädagogische Entscheidungen liefert (wie etwa zur Klassifikation von Schülern). Die Entwicklung des C-Tests ist der Versuch, ein globales Testverfahren zu verbessern, das bereits eine befriedigende theoretische (linguistische) Grundlage besitzt (general language proficiency / pragmatic expectancy grammar), aber methodische und testtheoretische Mängel aufweist.

Die beschriebenen Untersuchungsergebnisse verdeutlichen, daß der C-Test ein ökonomisches Verfahren ist, das sehr zuverlässig mißt. Dies gilt sowohl für L1-Lerner als auch für LF- und L2-Lerner. Der C-Test ist darüberhinaus sehr valide. Er mißt annähernd dasselbe, was auch in einer globalen Schulnote zum Ausdruck kommt. Seine Übereinstimmung mit Sprachtestbatterien ist sehr hoch. Dabei zeigt sich, daß der C-Test besonders hoch auf einem allgemeinen Sprachfaktor lädt und sich als globaler Sprachtest bestätigt. C-Tests besitzen nach ersten Untersuchungen in der Muttersprache eine mittlere Korrelation zu Intelligenztests und nur eine äußerst niedrige Korrelation zu Konzentrationstests (vgl. auch OLLER, 1978).

Trotz der sehr umfangreichen Untersuchungen mit C-Tests bestehen aber auch noch eine Reihe von Problemen. So ist die Untersuchung zur divergenten Validität und zur faktoriellen Struktur von C-Tests erst an einer kleineren Stichprobe analysiert worden. Weitere Untersuchungen, die verschiedene Klassenstufen einbeziehen und über den L1-Bereich hinausgehen, müssen sich anschließen. Untersuchungen zur prognostischen Validität von C-Tests bei unterschiedlichen Zielgruppen stehen noch aus. C-Tests sind außerdem bisher im Zweitsprachenbereich fast ausschließlich in der Grundschule eingesetzt worden, im Fremdsprachenbereich dagegen vorwiegend bei erwachsenen Sprachlernern. Für diese Zielgruppen sollten in anderen Altersgruppen weitere C-Tests entwickelt und erprobt werden. Das Problem der Textauswahl wurde bereits angeschnitten, es ist zur Zeit noch nicht zur Zufriedenheit gelöst. Weitere Kriterien sollten erarbeitet werden. Die Vorabschätzung der Textschwierigkeit scheint dagegen in naher Zukunft möglich zu sein. Dies würde die Testkonstruktion sehr erleichtern.

Die Übertragung des C-Prinzips auf andere Sprachen ist ebenfalls angesprochen worden. Bisher liegen Untersuchungen für deutsche, englische, französische, spanische und hebräische C-Tests vor (s. z.B. GROTJAHN & STEMMER, 1983; COHEN, SEGAL & WEISS, 1984). Die Entwicklung von C-Tests in anderen Sprachen wird z.T. in Kooperation mit ausländischen Kollegen durchgeführt. Insbesondere für Untersuchungen mit ausländischen Schülern in der Bundesrepublik erscheint es interessant, muttersprachliche Tests für die verschiedenen Nationalitäten zu konstruieren, so daß Sprachstandsvergleiche möglich werden oder aber Muttersprache und Zweitsprache im Rahmen anderer Forschungsfragen, etwa der Erforschung der Sprachlernbedingungen, reliabel und valide erfaßt werden können. Pilotstudien hierzu sind in Vorbereitung.

Wie die Entwicklung von deutschen C-Tests für ausländische Schüler gezeigt hat, kann das C-Prinzip selbst abgewandelt werden, um etwa leichtere Texte zu erhalten. Weitere Modifikationen sind denkbar und eventuell für bestimmte Stichproben oder auch für andere Sprachen angebracht. Auch sollte geprüft werden, ob zweideutige Lücken ebenfalls durch Abänderung des C-Prinzips entschärft werden können, etwa dadurch, daß ein Buchstabe mehr stehenbleibt. Möglicherweise können solche Probleme auch durch Wortersetzungen gelöst werden.

Die Erforschung der Anwendungsfelder von C-Tests hat gerade erst begonnen. C-Tests sind keine diagnostischen Tests; sie geben keine unmittelbare Information über Stärken und Schwächen eines Sprachlerners. Dazu müßten konventionelle, diagnostische Verfahren eingesetzt werden. C-Tests liefern aber globale Aussagen über den Sprachstand von Schülern, sowohl in der Muttersprache als auch in Fremd- und Zweitsprachen. Sie sind als Einstufungstests z.B. in Hochschulen sowie allgemein in Schulen geeignet. Ein interessantes Anwendungsfeld ist insbesondere ihr Einsatz bei ausländischen Schülern sowie zur Feststellung von Lernstörungen bei deutschen Schülern.

LITERATUR

Alderson, J.C.: The cloze procedure and proficiency in English as a foreign language. In: TESOL Quarterly, 1979, 13, 219-226.

Binet, A.: A propos de la mesure de l'intelligence. In: Année Psychologique, 1905, 11, 69-82.

Born, J.: Entwicklung und empirische Überprüfung eines Sprachstandstests für türkische Schüler der 2. Klasse. Praktikumsbericht, Erziehungswissenschaftliches Institut der Universität Düsseldorf. Düsseldorf, Januar 1984.

Bortz, J.: Lehrbuch der Statistik. Für Sozialwissenschaftler. Berlin: Springer, 1977.

Brown, J.D.: A correlational study of four methods for scoring cloze tests. Paper presented at the TESOL Convention. Boston, 1979.

Carroll, J.B.: Language development in children. In: Saporta, S. & Bastian, J.R. (eds.): Psycholinguistics: a book of readings. New York: Holt, Rinehart, and Winston, 1961, 331-345.

Carroll, J.B.: The psychology of language testing. In: Davies, A. (ed.): Language testing symposium. A psycholinguistic approach. London: Oxford University Press, 1968, 46-69.

Carroll, J.B., Carton, A.S. & Wilds, C.P.: An investigation of cloze items in the measurement of achievement in foreign languages. Research and Development Report, Laboratory for Research in Instruction. Harvard University, 1959.

Chomsky, N.: Syntactic structures. The Hague: Mouton, 1957.

Chomsky, N.: Aspects of the theory of syntax. Cambridge, Mass.: M.I.T. Press, 1965.

Cohen, A.C., Segal, M. & Weiss, R.: The C-Test in Hebrew. Unpublished manuscript. The Hebrew University of Jerusalem. Jerusalem, February 1984.

Davies, A.: Tests für den fremdsprachlichen Unterricht. In: Schrand, H. (Hrsg.): Testen. Probleme der objektiven Leistungsmessung im neusprachlichen Unterricht. Berlin: Cornelsen, Velhagen & Klasing, 1973, 23-44.

Davies, A.: The construction of language tests. In. Allen, J.P.B. & Davies, A. (eds.): Testing and experimental methods. London: Oxford University Press, 1977, 38-104.

Ebbinghaus, H.: Über die neue Methode zur Prüfung geistiger Fähigkeiten und ihre Anwendung bei Schulkindern. In: Zeitschrift für Psychologie, 1897, 13, 401-459.

Grimm, H.: Sprache. In: Klauer, K.J. (Hrsg.): Handbuch der Pädagogischen Diagnostik, Band 2. Düsseldorf: Schwann, 1978, 355-366.

Grotjahn, R. & Stemmer, B.: Entwicklung und Einsatz eines C-Tests 'Französisch'. Manuskript zu einem Vortrag auf der 14. Jahrestagung der Gesellschaft für Angewandte Linguistik (GAL) vom 29.9. bis 1.10.1983 an der Universität Duisburg. Duisburg, September 1983.

Hundsnurscher, F.: Sprachvielfalt, Sprachdynamik. In: Funk-Kolleg Sprache. Eine Einführung in die moderne Linguistik, Band 1. Frankfurt: Fischer Taschenbuch, 1973, 66-73.

Hymes, D.: Language in culture and society. A reader in linguistics and anthropology. New York: Harper and Row, 1964.

Ingenkamp, K. (Hrsg.): Die Fragwürdigkeit der Zensurengebung. Texte und Untersuchungsberichte. Weinheim: Beltz, 1977 (7. Auflage).

Jonz, J.: Improving on the basic egg: the multiple choice cloze test. In: Language Learning, 1976, 26, 255-265.

Klein-Braley, C.: Empirical investigations of cloze tests. An examination of the validity of cloze tests as tests of general language proficiency in English for German university students. Inauguraldissertation, Universität Duisburg, 1981.

Klein-Braley, C.: Textverständlichkeitsforschung. Manuskript zu einem Referat auf dem Ersten Kolloquium zur Quantitativen Linguistik an der Universität Essen. Essen, Februar 1983.

Klein-Braley, C.: Advance prediction of difficulty with C-Tests. To appear in: Culhane, T., Klein-Braley, C. & Stevenson, D.K. (eds.): Practice and problems in language testing 7. Proceedings of the Seventh International Language Testing Symposium of the Interuniversitäre Sprachtestgruppe (IUS). Colchester: University of Essex, 1984.

Klein-Braley, C. & Lück, H.E.: Entwicklung des Duisburger Englisch-Leistungstests für Anglistikstudenten (DELTA). Bericht Nummer 2 aus dem Arbeitsbereich Psychologie der Fernuniversität Hagen. Hagen, April 1979.

Lado, R.: Language testing. The construction and use of foreign language tests. A teacher's book. London: Longman, 1961. Deutsche Fassung: Testen im Sprachunterricht. München: Hueber, 1971.

Lienert, G.A.: Testaufbau und Testanalyse. Weinheim: Beltz, 1969 (3. Auflage).

Oller, J.W., Jr.: Cloze tests of second language proficiency and what they measure. In: Language Learning, 1973, 23, 105-118.

Oller, J.W., Jr.: How important is language proficiency to IQ and other educational tests? In: Oller, J.W., Jr. & Perkins, K. (eds.): Language in education: testing the tests. Rowley, Mass.: Newbury House, 1978, 1-15.

Oller, J.W., Jr.: Language tests at school: a pragmatic approach. London: Longman, 1979.

Oller, J.W., Jr. & Conrad, C.: The cloze technique and ESL proficiency. In: Language Learning, 1971, 21, 183-195.

Oller, J.W., Jr. & Perkins, K. (eds.): Research in language testing. Rowley, Mass.: Newbury House, 1980.

Preibusch, W., Hagemeister, V., Schuricht, K. & Seyhan, H.: Cloze-Tests als Instrumente zur Kontrolle der Deutsch-Kenntnisse türkischer Schüler in der Sekundarstufe I. In: Unterrichtswissenschaft, 1983, 11, 180-193.

Raatz, U.: Sprachtheorie und Faktorenanalyse. IUS-Workpapers 9. Essen und Duisburg: Interuniversitäre Sprachtestgruppe, November 1981.

Raatz, U.: The factorial validity of C-Tests. To appear in: Culhane, T., Klein-Braley, C. & Stevenson, D.K. (eds.): Practice and problems in language testing 7. Proceedings of the Seventh International Language Testing Symposium of the Interuniversitäre Sprachtestgruppe (IUS). Colchester: University of Essex, 1984.

Raatz, U. & Klein-Braley, C.: The C-Test: a modification of the cloze procedure. In: Culhane, T., Klein-Braley, C. & Stevenson, D.K. (eds.): Practice and problems in language testing 4. Proceedings of the Fourth International Language Testing Symposium of the Interuniversitäre Sprachtestgruppe (IUS). University of Essex, Occasional papers 26. Colchester: University of Essex Press, 1982, 113-138.

Raatz, U. & Klein-Braley, C.: Ein neuer Ansatz zur Messung der Sprachleistung. Der C-Test: Theorie und Praxis. In: Horn, R., Ingenkamp, K. & Jäger, R.S. (Hrsg.): Tests und Trends 3. Jahrbuch der Pädagogischen Diagnostik. Weinheim: Beltz, 1983, 107-138.

Raatz, U., Klein-Braley, C. & Süßmlich, E.: Der C-Test: Ein Verfahren zur allgemeinen Sprachstandsmessung im mutter- und fremdsprachlichen Unterricht. Manuskript zu einem Referat auf dem 12. Kongreß für angewandte Psychologie vom 21. bis 25. September 1983 an der Universität Düsseldorf. Düsseldorf, September 1983.

Salimbene, S.: From structurally based to functionally based approaches to language testing: a movement from teacher centered to students centered philosophies of language education. Paper presented at the Fourteenth Annual TESOL Convention. San Francisco, 1980.

Sang, F. & Vollmer, H.J.: Allgemeine Sprachfähigkeit und Fremdsprachenerwerb. Zur Struktur von Leistungsdimensionen und linguistischer Kompetenz des Fremdsprachenerwerbs. Diskussionsbeiträge aus dem

Institut für Bildungsforschung, Heft 1. Berlin: Max-Planck-Institut für Bildungsforschung, 1978.

Schröter, G.: Die ungerechte Aufsatzzensur. Bochum: Kamp, 1971.

Spolsky, B.: What does it mean to know a language, or how to get someone to perform his competence? Paper presented at the Second Conference on Problems in Foreign Language Testing. University of Southern California, November 1968.

Spolsky, B.: Language testing: art or science? Main lecture delivered at AILA World Congress '75. University of Stuttgart, August 1975.

Spolsky, B.: Some ethical questions about language testing. In: Klein-Braley, C. & Stevenson, D.K. (eds.): Practice and problems in language testing 1. Proceedings of the First International Language Testing Symposium of the Interuniversitäre Sprachtestgruppe held at the Bundessprachenamt, Hürth, 29-31 July 1979. Frankfurt: Lang, 1981, 5-30.

Spolsky, B., Sigurd, B., Sato, M., Walker, E. & Arterburn, C.: Preliminary studies in the development of techniques for testing overall second language proficiency. In: Upshur, J.A. & Fata, J. (eds.): Problems in foreign language testing. Language Learning, Special Issue Number 3, 1968, 79-101.

Steinert, J.: Allgemeiner Deutscher Sprachtest. Handanweisung für die Durchführung, Auswertung und Interpretation. Braunschweig: Westermann und Göttingen: Hogrefe, 1978.

Stevenson, D.K.: The sociolinguistics of language testing: a tester's perspective. In: Lutjeharms, M. & Culhane, T. (eds.): Practice and problems in language testing 3. Proceedings of the Third International Language Testing Symposium of the Interuniversitäre Sprachtestgruppe (IUS). Brussel: Vrije Universiteit Brussel, 1982, 4-22.

Süßmilch, E.: Sprachdiagnostik bei Ausländerkindern. In: Ingenkamp, K. (Hrsg.): Sozial-emotionales Verhalten in Lehr- und Lernsituationen. Bericht über die 34. Tagung der Arbeitsgruppe für empirische pädagogische Forschung in der DGfE vom 28. bis 30. September 1983 in Landau/Pfalz. Erziehungswissenschaftliche Hochschule Rheinland-Pfalz, Landau 1984a, 375-377.

Süßmilch, E.: Language testing with immigrant children. To appear in: Culhane, T., Klein-Braley, C. & Stevenson, D.K. (eds.): Practice and problems in language testing 7. Proceedings of the Seventh International Language Testing Symposium of the Interuniversitäre Sprachtestgruppe (IUS). Colchester: University of Essex, 1984b.

Süßmilch, E. & Raatz, U.: Entwicklung und Erprobung eines Tests zur Erfassung der allgemeinen Sprachleistung ausländischer Grundschüler. Manuskript zu einem Vortrag auf der 2. Tagung der Regionalen Arbeitsstellen (RAA) in Essen am 26. November 1982 zum Thema: Sprachstandsmessung bei ausländischen Kindern und Jugendlichen. Essen, November 1982.

Süßmilch, E. & Raatz, U.: Neuere Möglichkeiten der Sprachstandsmessung bei Ausländerkindern. Manuskript zu einem Vortrag auf der 14. Jahrestagung der Gesellschaft für Angewandte Linguistik (GAL) vom 29.9. bis 1.10.1983 an der Universität Duisburg. Duisburg, September 1983.

Taylor, W.L.: Cloze procedure: a new tool for measuring readability. In: Journalism Quarterly, 1953, 30, 415-433.

Taylor, W.L.: Recent developments in the use of cloze procedure. In: Journalism Quarterly, 1956, 33, 42-48.

Ulrich Esser
Karl-Marx-Universität
Leipzig, DDR

PSYCHOLOGISCHE ASPEKTE DER DIAGNOSE VON FREMDSPRACHENLERNFÄHIGKEITEN - EINE DIAGNOSE DER DIAGNOSTIK

Die Diagnose dessen, was ein Fremdsprachenlerner an geistigen Voraussetzungen mitbringen muß, um eine Fremdsprache (FS) zu lernen, was er gelernt hat im Verlaufe eines Sprachkurses und wie er in der Lage ist, mit diesem Wissen kommunikative Probleme in der FS lösen zu können, ist nicht nur legitim, sondern in mehrfacher Hinsicht unbedingt notwendig (Hellmich, Desselmann 1981). In den letzten Jahren haben sich in der Folge dieser Einsicht die Bemühungen einer sog. Fremdsprachendiagnose explosionsartig verstärkt, so daß die Zahl der gegenwärtig angebotenen kommerziellen Testverfahren kaum noch überschaubar ist. Hoffmann (1962) spricht von ca. 8 000 solcher Tests, Jones (1977) schätzt die Zahl bereits auf 12 000 bis 13 000 Eignungs- und Leistungstests. Dieses hauptsächlich pragmatistisch betriebene Vorgehen in der Testkonstruktion der fremdsprachigen Lehrpraxis, verbunden mit den an solche Diagnosen verknüpften Entscheidungen über zukünftige Entwicklungen der Untersuchten (Berufslaufbahn, Studienentscheidungen etc.) führte zu einem allgemeinen Unbehagen in der Fremdsprachendiagnostik. Die theoretischen Grundlagen des diagnostischen Vorgehens, die Art und Weise des technischen Vorgehens als auch die Zielsetzungen werden immer mehr und mehr in Frage gestellt und eine Neuorientierung der Fremdspracheneignungs- und leistungsdiagnose wird angestrebt.

Versucht man schlagwortartig die gegenwärtige Situation der FS-Diagnostik zu kennzeichnen, so dominieren folgende diagnostischen Zugänge:

- personenzentrierte Diagnose: im Sinne der Erfassung streng isolierbarer invarianter Eigenschaften und Merkmale von Lernern, wie spezielle Dimensionen sprachlicher Fähigkeiten, sprachlicher Fertigkeiten, Wissensbeständen etc. Die noch aus der Harris-Valette-Ära stammenden "discrete-point-tests vs. integrated tests" sind die besten Beispiele für eine solche personenzentrierte Diagnose.
- statusorientierte, konstatierende Ein-Punkt -Messung: diagnostiziert

wird die Ausprägung bzw. das Vorhandensein vs. Nichtvorhandensein bestimmter Merkmale zu einem ganz bestimmten, fest definierten Zeitpunkt. Der dynamische Charakter dieser Merkmale bzw. Prozesse, ihre Veränderbarkeit durch die Messung selbst bzw. im Verlaufe selbst kurzer Zeitintervalle wird weitgehend aus der Diagnose ausgeklammert.
Ein solcher diagnostischer Zugang erlaubt lediglich Hier- und -Jetzt - Entscheidungen von diagnostischem, konstatierendem Wert. Irgendwelche Prognosen über zukünftiges Verhalten, über die künftige Entwicklung sind weitgehend ausgeschlossen oder nur mit großen Problemen verbunden. Aufgabe des Diagnostikers ist es aber hauptsächlich, Prognosen zu treffen.

- Effekt- oder ergebnisorientierte Diagnose: Grundlage für diagnostische Entscheidungen im FS-Bereich sind immer noch die Ergebnisse richtig oder falsch gelöster Testausgaben, ohne der Frage nachzugehen, auf Grund welcher kognitiven/geistigen Prozesse und Leistungen diese Ergebnisse entstanden sind. Das Vorgehen erinnert stark an die Formen eines behavioristisch geprägten Positivismus in der Leistungsdiagnose: Input ist die Testaufgabe, Output ist das gelöste Ergebnis und lediglich aus der Beziehung zwischen In- und Output wird versucht, auf die inneren Mechanismen zu schließen. Der Lösungsmechanismus selbst wird als eine Art black-box betrachtet. Fakt ist aber, daß ein und dasselbe Resultat auf Grund unterschiedlichster Lösungswege und Lösungsformen entstanden sein kann. Und die zu erfragen, sollte im Vordergrund der Diagnose stehen, um zu hinreichend validen Aussagen zu gelangen.
- Stichprobenorientierte Diagnose: immer noch werden die Versuchspersonen (Vpn) entsprechend ihrer Testpunktwerte an den Leistungsnormen entsprechender Stichproben von Personen klassifiziert und nicht entsprechend anforderungsorientierter Leistungsmaßstäbe.
- Diagnose im Stile labormäßiger Prüfungssituationen: bedingt durch den Prüfungscharakter und durch eine gewisse "weltfremdheit" der Tests wirft sich die Frage auf, ob das, was mit den Tests getestet wird auch das ist, was in der realen Sprachverwendungssituation den Wert hat, den man ihm a priori zuschreibt. Gerade an diesem Punkt schließt sich die hoch aktuelle Diskussion um die ökologische Validität psychologischer bzw. allgemeiner: sozialwissenschaftlicher Untersuchungen an. Solche momentan kritischen Fragen wie die einer unangemessenen Wahl zugrundeliegender Meßmodelle für die Konstruktion von FS-Prüfverfahren etc. seien im Rahmen dieses Beitrages ausgeklammert.

Mit der Auflistung dieser Merkmale sind im wesentlichen auch die Perspektiven der Entwicklung der fremdsprachigen Eignungs- und Leistungsdiagnose in der Zukunft gekennzeichnet. Sie liegen in einer dialektischen Negation der "Krisenmerkmale", ohne aber dabei die Relevanz dieser diagnostischen Zugänge in Frage stellen zu wollen:

- lernzuwachsorientierte Diagnose
- veränderungs- und prozeßorientierte Diagnose
- tätigkeitsorientierte Diagnose
- anforderungsorientierte Diagnose, verbunden mit der Nutzung nicht-konventioneller Meßkalküle
- ökologisch valide Diagnose.

Im folgenden sei versucht, einige Grundgedanken dieser methodologisch neuorientierten Richtung in der FS-Diagnose am Beispiel der FS-Lernfähigkeits: diagnose zu exemplifizieren.

FS-diagnostische Unternehmungen sind im wesentlichen auf die Erfassung dreier zentraler psychischer Konstrukte gerichtet: der FS-Lernfähigkeit, FS Wissens- und Kenntnisbestände, des FS Könnens. Die Auffassungen über den Bedeutungsgehalt dieser Begriffe und ihrer gegenseitigen Beziehungen sing allgemein sehr uneinheitlich und teilweise kontrovers (Hellmich, Desselmann 1981, Esser 1983). Aber gerade im Verständnis dieser Begriffe lassen sich Gegenwart und Zukunft der FS Lernfähigkeitsdiagnose wenigstens skizzenhaft andeuten.

Der Erwerb einer FS (egal, ob unter pädagogisch gesteuerten Bedingungen oder ungesteuert) setzt bestimmte geistige Voraussetzungen als notwendig voraus, die wesentlich das Niveau und die Qualität des FS Lernsprozesses und seiner Lernergebnisse bestimmen und damit dem Lerner ein kommunikativ kompetentes Verhalten in mannigfaltigen Problemsituationen erlauben. Diese Menge an geistigen Voraussetzungen, die den Erwerb einer FS, aber noch nicht ihren Gebrauch bestimmen, macht die *Fremdsprachenlernfähigkeit* im Sinne einer Lernpotenz bzw. der Lernmöglichkeit aus. Die FS Lernfähigkeit wiederum ist Bestandteil einer umfassenderen FS *Lernsdisposition* die zusätzlich noch Voraussetzungen für den erfolgreichen Gebrauch einer FS enthält, also motorische und anatomisch-physiologische Faktoren.

Die Auffassungen über die Existenz, die Struktur und die Funktion des psychischen Konstruktes der FS Lernfähigkeit oder FS-Eignung sind nicht einheitlich, bedingt durch eine gewisse "Theorienabstinenz" (Esser 1982) in der Wahl der Argumente und in der Anlage der Experimente. Einerseits wird die lernbeeinflussende Funktion eines solchen psychischen Konstruktes weitgehend abgestritten, andererseits raäumen Autoren wie Gardner, Lambert (1965), Dale (1976) der FS- Lernfähigkeit zwar eine gewisse lernbeeinflussende Funktion beim FS- Erwerb ein, aber - so nach Dale - und in gewisser Weise hat er sogar recht -: "Wenn die soziale Situation es verlangt, dann lernt jeder eine Fremdsprache zu beherrschen, unabhängig von seinen Fähigkeiten und seiner Intelligenz".

Beim Konzept der FS-Lernfähigkeit als einem zwar real existierenden aber nicht unmittelbar beobachtbaren Konstrukt wird stillschweigend unterstellt, daß die kognitiven Dimensionen zielsprachenunabhängig und bereits vor dem ersten Kontakt mit der FS existent sind. Zu welchen Anteilen jedoch in die FS Lernfähigkeit allgemeine (universale) und damit generell sprachunabhängige kognitive Voraussetzungen, im Prozeß des Mutterspracherwerbs aufgebaute (und dann reaktivierte) Lernmechnaismen und anlagebedingte Mechnaismen mit eingehen, ist weitgehend offen.

Bspw. konnten Gerver, Esser (1976) zeigen, daß deutsche Muttersprachler mit einem hohen formalen Intelligenzfaktor eine flektierte Sprache wie Russisch schneller und leichter lernten als Englisch. Lerner mit einem hohen mechanischen Gedächtnisfaktor lernten dagegen Englisch schneller als Russisch. Vermuten läßt sich, daß spezielle kognitive Leistungsvoraussetzungen den FS-Erwerb beeinflussen und daß diese Faktoren durchaus nicht zielsprachenunabhängig sein müssen.

Unbestreitbar ist, daß das Konstrukt der Fremdsprachenlernfähigkeit im Prozeß der aktiven muttersprachlichen Tätigkeit (eingebettet in den allgemeinen Prozeß der Lebenstätigkeit) und unter Einfluß bestimmter historischer, sozialer und personaler Bedingungen entwickelt wird und sich im Prozeß eines weiteren Spracherwerbs weiterverändern wird. Es ist damit ein dynamisches, entwicklungsabhängiges Konzept. Insofern wird die stillschweigende Annahme: "es ist bereits vor dem ersten Kontakt mit der FS existent" problematisiert. In eigenen Untersuchungen zum Kunstsprachenerwerb (Esser 1980) ließ sich bspw. zeigen, daß bestimmte Dimensionen der FS-Lernfähigkeit durch den Erwerb von Miniaturkunstsprachen trainierbar sind derart, daß sich der nachfolgende Erwerb einer natürlichen FS auf eben dieser angestrebten Dimension erleichtert.

Welche konkreten Dimensionen und kognitiven Voraussetzungen die FS Lernfähigkeit ausmachen, ist gegenwärtig noch weitgehend unklar. Als sicher kann jedoch angenommen werden, daß das Niveau und die Qualität des FS-

Lernprozesses u.a. durch solche Fähigkeiten bestimmt wird wie:
- "linguistic reasoning" als ein allgemeiner Intelligenzfaktor,
- Wortflüssigkeit in der Muttersprache,
- Fähigkeit zur Einsicht in die Struktur einer Sprache und die Fähigkeit, grammatische Zusammenhänge zu erfassen,
- Fähigkeit des (mechanischen und sinnhaften) Behaltens von Paarassoziationen,
- induktive Sprachlernfähigkeit (oder Fähigkeit des induktiven Regelwerbs),
- phonetische Kodierungsfähigkeit etc.

Man könnte die Liste weiter fortsetzen.

Alle Fähigkeitsdimensionen sind dabei ausschließlich faktorenanalytisch nach der Konstruktion und dem Einsatz sog. FS-Fähigkeitstests ermittelt worden und nicht der Testkonstruktion bereits zugrundegelegt worden. Damit ist keinesfalls ausgeschlossen, daß bisher eine Menge von solchen kognitiven Voraussetzungen erfaßt wurde, die zwar für die Bearbeitung der speziellen Tests notwendig ist, aber nicht in gleichem Maße auch notwendig ist für den realen Erwerb einer solchen Fremdsprache. Versuche, aus den Testergebnissen bisheriger Eignungstests *Prognosen* über den zukünftigen Lernerfolg abzuleiten, haben sich als wenig erfolgreich herausgestellt. Dieser Mißstand muß aber nicht nur auf der Tatsache beruhen, daß eine Prognose auf der Grundlage von Statustests kaum möglich ist, sondern kann auch dadurch bedingt sein, daß die erhaltenen FS-Fähigkeitsdimensionen keine unbedingte reale Lernrelevanz besitzen. In Anlehnung an die sarkastische Bemerkung in der Intelligenzdiagnostik könnte man auch für die FS-Fähigkeitsdiagnostik sagen: FS-Lernfähigkeit ist das, was FS-Lernfähigkeitstests messen!

Der gegenwärtige Zustand in der Diagnose der FS-Lernfähigkeit und der damit verbundenen Theorienbildung ist nicht befriedigend. Zwei Gründe scheinen dafür verantwortlich zu sein:

(1) Das bisherige Meßkonzept der FS-Lernfähigkeitsdiagnose trägt weitgehend den Charakter einer konstatierenden Statusmessung. Gemessen wird das, was momentan existent ist, was vor dem ersten Kontakt mit der FS vorliegt.

Diesem Konzept der konstatierenden Statusmessung liegt die in der älteren Psychologie vorherrschende Auffassung zugrunde, wonach das, was ein Mensch in der Vergangenheit gelernt hat (sein gegenwärtiger Entwicklungsstand) auch ein gültiger Indikator für das sei, was er in Zukunft lernen könne. Eine Auffassung, die bereits in Arbeiten von Rubinstein, Wygotski etc. als überholt und unzutreffend dargestellt wurde.

(2) Man ist bisher bei der Ermittlung der kognitiven Voraussetzungen für den FS-Erwerb zu sprachspezifisch bzw. zu sprachnahe herangegangen. Sowohl in zahlreichen allgemeinspsychologischen Experiemnten (Neisser 1974) als auch in mehreren eigenen Untersuchungen (Esser 1980, 1982) ließ sich zeigen, daß die Wirksamkeit bisheriger Dimensionen der FS-Lernfähigkeit eine Funktion allgemeinerer kognitiver Voraussetzungen wie z.B. denen des induktiven Regelerwerbs, der Analogiebildung, der allgemeinen Diskriminationsfähigkeit ist. Ein ernstzunehmender diagnostischer Ansatz hätte bei diesen Faktoren anzusetzen.

Zu Punkt 1: Nahezu alle bisherigen FS-Eignungstests zeichnen sich dadurch aus, daß an einem Probanden vor Beginn der FS-Lernphase eine einmalige Testerhebung vorgenommen wird, auf deren Grundlage eine, wenn auch worsichtige Prognose über den zukünftigen Lernerfolg geteroffen wird. Diese prognostische Validität (im Sinne einer Voraussagefähigkeit) der einschlägigen FS-Eignungstests liegt zwischen 0,3 und 0,5 und damit nicht sehr hoch. Eine ausreichend gesicherte Prognose zukünftiger Lernleistungen ist auf der Grundlage der bisherigen diagnostischen Vorgehensweisen nur mit großer Vorsicht möglich.

Kognitive Prozesse sind nicht unveränderlich. Entscheidend für den FS-Erwerb ist nicht nur das, was der Lernende an kognitiven Voraussetzungen mitbringt (der gegenwärtige Status bzw. Entwicklungsstand), sondern in welchem Maße er bei Existenz eines bestimmten kognitiven Status befähigt ist, eine Fremdsprache zu lernen. Erst ein Meßkonzept, das sowohl die Statusmessung als auch die Messung einer solchen Lernfähigkeit in sich vereint, wird in der Lage sein, die gestellten Forderungen an FS-Lernerfolgsprognosen hinreichend zu erfüllen.

Die dialektische Negierung der bisherigen traditionellen Statusmessung, und damit eine der relevanten Perspektiven der FS-Lernfähigkeitsdiagnose, basiert auf der von Wygotski (1934/1964) konzipierten Idee der "Zone der nächsten Entwicklung". Nach Wygotski hat nicht mehr die reine Konstatierung des augenblicklichen Entwicklungsstandes der geistigen Fähigkeiten im Vordergrund zu stehen. Vielmehr muß der reale (unter pädagogischen Einwirkungen vonstattengehende) Lernprozeß simuliert bzw. modelliert werden. Durch diesen in der diagnostischen Situation provozierten Lernprozeß wird die "Zone der nächsten Entwicklung" sichtbar, und damit die eigentliche

Lernpotenz im Sinne zukünftiger Lernmöglichkeiten bestimmbar. Die operationale Konsequenz dieser Idee war die Entwicklung sog. Lerntests, in denen das Prinzip der Pädagogisierung der Fähigkeitsdiagnostik (wie es besonders in der sowjetischen Psychologie entwickelt wurde; Itelson 1967, Kalmykowa 1975, Leontjev 1975, s.a. Guthke 1978) mid dem hochentwickelten Verfahrens- und Auswertungstechniken des psychometrischen Ansatzes verbunden werden kann und die eine Abbildung des Sprachlernprozesses im "kleinen Stil" darstellen. Der Grundaufbau solcher Lerntests ist - bei aller Variation - wie folgt: Pretestung - Pädagogisierungsphase: in Form von Sanktionen, pädagogischen Hilfen etc. - Poststestung. Allgemeine Anforderungen an die Entwicklung der Verfahren und an die Formulierung der Testaufgaben sind u.a. (s.a. Harnisch 1983):

- Homogenität des Testmaterials
- Möglichkeit einer vollständigen Analyse des Lösungsprozesses
- stetige Zunahme der Aufgabenkomplexität derart, daß die richtige Lösung der Aufgabe Nr. i die Voraussetzung für die erfolgreiche Lösung der Aufgabe Nr. iß1 ist
- Möglichkeit einer individuellen, standardisierten Pädagogisierung in Form einer positiven/negativen Rückkopplung, von Hilfen etc.

Ansätze einer lernzuwachsorientierten Diagnose der FS-Lernfähigkeit sind gegenwärtig nicht bekannt, obwohl viele existente Verfahren die Möglichkeit einer solchen Vorgehensweise besitzen. Zu Ansätzen im Bereich der Intelligenzdiagnostik: siehe Guthke (1978).

Die bisher einzigen erfolgversprechenden diagnostischen Zugänge zur Diagnose der FS-Lernfähigkeit wurden von Harnisch (1983), in teilweiser Zusammenarbeit mit dem Autor und von Esser (1980, 1982) entwickelt und unter verschiedensten Bedingungen erprobt. Diagnostiziert wurde die Fähigkeit des syntaktischen Regelerwebs unter Nutzung der kognitiven Operationen des logischen Schließens (als einer Operation, die dem linguistic reasoning unterliegt) und der Analogiebildung. Die Probanden hatten successiv die Konstruktionsregeln einer Miniaturkunstsprache (unterschiedlicher Konstruktionskomplexität) mit und ohne referentiellem Feld zu erlernen. Die diagnostische Aufgabenstellung lautete: "Figuren auf der einen Seite werden durch Sätze einer Kunstsprache auf der anderen Seite des Testblattes repräsentiert. Versuchen Sie dabei (durch logisches Schließen) die Lücken in den Sätzen auszufüllen!" Die Aufgaben wurden so eingeführt, daß durch die Analyse konstant bleibender Merkmale der Ikone und Sätze je ein variierendes Merkmal erschlossen werden kann, das für die Lösung nachfolgender Schlüsse notwendig war. Die Pädagogisierung erfolgte als positive oder negative Rückmeldung bzw. als gestaffelte Zusatzinformation.

Aus der Zahl der Lernschritte, der Fehler sowie der notwendigen Hilfen sollte auf die FS-Lernfähigkeit im Sinne der "Zone der nächsten Entwicklung" geschlossen werden.

Eine typische Beispielaufgabe der Tests sah wie folgt aus:

□ ski

□ ○ ski la gadu

○ ski vep gadu
□

○ □ ___ ___ ___

In dieser Lerntestvariante werden Aspekte des Erwerbs einer natürlichen Fremdsprache durch den Erwerb einer Miniaturkunsstsprache simuliert. Die Auswertung der Ergebnisse zeigte eine gegenüber anderen FS-Lernfähig-keits-tests signifikant höhere prognostische Validität in verschiedenen Altersstufen (6. Klasse, 9. Klasse, Universität), verschiedenen Sprachen (Russisch, Englisch, Deutsch) und verschiedenen Ausbildungsformen (u.a. auch Intensivkurse). Offensichtlich werden mit den Lerntestvarianten von Harnisch und Esser u.a. universale, sprachinvariante kognitive Voraussetzungen des FS-Erwerbs diagnostiziert. Kriterien für die prognostische Validierung waren die Noten der Lernenden im Sprechen und Verstehenden Hören. Der Zeitraum zwischen Testeinsatz und Erhebung der Lernerfolgskriterien betrug 3 Monate (+/- 14 Tagen). Die Korrelation der Testpunktwerte mit den Noten ergab folgende Ergebnisse:

	$r_{tc/progn.}$
Harnisch-Test	0,62
Esser-Kunstsprachentest	0,55
York-FS-Eignungstest (Green)	0,46
dtsch. Fassung des MLAT (Caroll, Sapon)	0,49
Cattell-g-Faktor-Test	0,31

Eine Interpretation der Ergebnisse erübrigt sich.

Auch wenn die gegebene Skizze möglicher Perspektiven der FS-Diagnostik nur sehr grob erfolgen konnte, so sei doch der untrennbare Zusammenhang zwischen Theorie, Methodologie und Operationalisierung in Form von diagnostischen Prüfverfahren etwas deutlicher herausgearbeitet worden am Beispiel der Lernfähigkeitsdiagnose. Abschließend noch zwei Bemerkungen: Diagnose und Messung eines Sachverhaltens sind durchaus nicht gleichzusetzen. Nicht nur, daß die bisherigen statischen und konstatierenden Formen der Ergebniserfassung zugunsten einer prozeß- und tätigkeitsorientierten Diagnose aufzugeben sind. Notwendig ist es auch in der Perspektive, neben den rein psychometrischen Vorgehen stärker als bisher außerintellektuelle, speziell motivationale Faktoren der FS-Lernfähigkeit in die Diagnose einzubeziehen. Erst eine solche breiter angelegte "Diagnostik der Lernerpersönlichkeit" wird den vielfältigen und steigenden sozialen Anforderungen besser gerecht werden können als die bisher praktizierten, etwas einseitigen psychometrischen Versuche der FS-Lernfähigkeitsdiagnose.

Ebenso ist einsichtig, daß durch das Problem der FS-Eignungsdiagnostik das Problem der Optimierung des FS-Erwerbs nicht gelöst werden kann, ebensowenig, wie durch immer neuere methodische und didaktische Varianten der Vermittlung in ein und demselben Vermittlungsparadigma. Solange nicht bekannt ist, wie Lernende predisponiert sind, welche individuellen Lernstile optimal sind, welche Sprachen welche Lehrmethoden verlangen, werden alle Optimierungsbemühungen von vornherein weitgehend eingeschränkt bleiben.

Es sei hier nicht für ein individualisiertes Herangehen im Fremdsprachenunterricht plädiert, jedoch für eine stärkere Differenzierung des bisher noch zu globalen und lehrerzentrierten Zuganges der FS-Vermittlung -- und auch Diagnose.

LITERATUR

Carroll, J.B.: Twenty-five years research on foreign language aptitude, North Carolina 1979

Esser, U.: Personale und soziale Faktoren des Fremdsprachenerwerbs, in: Deutsch als Fremdsprache, 2/1983, 71-77

Esser, U.: Ein allgemeinpsychologischer Ansatz zur Diagnose der FS-Lernfähigkeit, in: Deutsch als Fremdsprache, 2/1982, 76-82

Esser, U.: Miniaturkunstsprachen- ein möglicher Zugang zur Diagnose der FS-Lernfähigkeit, Karl-Marx-Universität Leipzig, unveröff. Forschungsbericht, 1980

Dale, P.S.: Language development: structure and function, New York 1976

Gardner, R.C., Lambert, W.E.: Language Aptitude, Intelligence and Second Language Achievement, in: Journal for Education and Psychology 56, 1965, 191-199

Gerver, D., Esser, U.: Intelligence and Foreign Language Acquisition, unpubl. Research Paper, University of Leipzig, Department of Psychology, 1976

Guthke, J.: Ist Intelligenz meßbar? Berlin: VdW 1978

Harnisch, A.: Die Entwicklung eines diagnostischen Programms zur Untersuchung von syntaktischem Regel- und Lexikerwerb, in: Esser, U., Hartl, B. (Hrsg.): Gegenwärtige Probleme und Aufgaben der Fremdsprachenpsychologie, Leipzig 1983

Hellmich, H., Desselmann, G.: Didaktik des Fremdsprachenunterrichts, Leipzig, Enzyklopäd.V. 1981

Hoffmann, G.: The tyranny of testing, New York 1962

Itelson, L.: Mathematische und kybernetische Methoden in der Pädagogik, Berlin 1967

Jones, R.L.: Testing: A vital connection, in: Phillips, J.K. (ed.) The Language Connection: From the Classroom to the World, Skokie, Ill. 1977

Jones, R.L.: An International Survey of Research in Language Testing: 1977-1979, Provo (Utah) 1980

Kalmykowa, S.J. (Hrsg.): Probleme der Diagnostik der geistigen Entwicklung der Schüler, Moskau (russ.) 1975

Leontjev, A.N.: Tätigkeit, Bewußtsein, Persönlichkeit, Moskau (russ.) 1975

Neisser, U.: Kognitive Psychologie, Stuttgart 1974

Pimsleur, P., Stockwell, R.P., Comrey, A.L.: Foreign Language Ability, in: Journal of Education and Psychology 53, 1962, 15-26

Vollmer, H.J.: Sprqcherwerb und Sprachbeherrschung, Tübingen, NarrV. 1982

Witzlack, G.: Grundlagen der Psychodiagnostik, Berlin: Verl.d.Wissenschaften 1977

Wygotski, L.S.: Denken und Sprechen, Berlin 1964 (russ. 1934)

Waldemar Marton
Institute of English, Adam Mickiewicz University
Poznań, Poland

THE CONTRASTIVE ANALYSIS HYPOTHESIS AND THE CONTEMPORARY ACQUISITIONAL PARADIGM: A COMPARISON WITH SOME PEDAGOGICAL IMPLICATIONS

As it is commonly known, the Contrastive Analysis Hypothesis was the belief that structural differences between the target language and the learner's native language are a major source of difficulty and error in the process of acquiring that target language and that we can predict both error and difficulty on the basis of contrastive analysis of the two languages involved. This belief was deeply rooted in structural linguistics and behavioristic psychology, which viewed language as a set of habits and language acquisition as habit formation. It is also known that along with the development of new thinking about language and in view of mounting criticism levelled against behavioristic linguistics the Contrastive Analysis Hypothesis has undergone various changes and modifications, from the original strong version to the weak version formulated by Wardhaugh (1970) and then to the moderate version suggested by Oller and Ziahosseiny (1970). In spite of these modifications, however, it can be safely said that the Contrastive Analysis Hypothesis has been quietly laid to rest by most contemporary researchers and language teaching theorists and that the claims made by it are no longer seriously considered by contemporary literature on language learning and teaching.

Yet it seems to me that the time has come to look at the hypothesis in question anew, from the perspective of present-day knowledge, and to compare it with the contemporary scientific paradigm to see whether the latter has greater explanatory power and greater practical value. This practical value refers, in my understanding, to the three important pedagogical processes of error prevention, error correction and error eradication. The discussion of all these issues constitutes the subject matter and the purpose of this paper.

Let me begin the discussion by stating that the above-mentioned contemporary paradigm, stemming from transformational-generative linguistics and first and second language acquisition research, holds that language is a set of rules and that language acquisition is basically the process of rule learning or, in other words, the process of hypothesis formation and testing. Accordingly, the influence of the native language, which is what the contrastive analysis hypothesis is all about, can be said to affect the process of hypothesis formation rather than the process of habit formation. This statement, in turn, provokes a rather obvious question, namely, whether this reformulation of the Contrastive Analysis Hypothesis is a mere verbal trick or whether it truly points to a new quality, to a new and better understanding of second language acquisition. I think that the latter is the case and I will try to defend this view in the remaining part of the paper.

First of all, to be fair in our evaluation of the Contrastive Analysis Hypothesis we should recognize the fact that, being a behavioristic notion, it was only marginally concerned with accounting for mental processes responsible for language acquisition. Its main interest was in finding observable causes of erroneous behavior and in describing the consequences of this behavior. Accordingly, it induced a blanket view of second language acquisition, in terms of which every structural divergence between L_1 and L_2 was seen as a most likely source of difficulty and error for every learner.

Today, on the contrary, we are primarily interested in discovering the nature of mental processes responsible for second language acquisition. Taking this fact into consideration we can say that the contemporary acquisitional paradigm can be seen as a complementation and elaboration of the Contrastive Analysis Hypothesis. The latter, on the other hand, can be viewed in retrospect as a fairly naive and simplistic theoretical overgeneralization.

What are, then, the consequences of the view that the native language influences the process of hypothesis formation and testing in second language acquisition? Of course, owing to second language acquisition research we know that the projection of the knowledge of L_1 is only one source for making hypotheses about L_2, the other and the essential source being observation of the linguistic data carried out in accordance with some, most probably innate, cognitive principles.

After making this reservation we may conclude that one obvious consequence of this view is that the influence of L_1 on the process of second language acquisition is a very individual thing, depending to a

large extent on the individual learner. This conclusion is based on the fact that it is well known in educational psychology that hypothesis formation depends on a whole lot of personal variables such as the learner's existing cognitive structure, his motivation, his aptitude, his preferred learning style, his preferred tactic of language acquisition (for the notion of language acquisition tactic, see Marton 1983), and some others. Since these variables are not necessarily stable and may change more or less quickly in the learner's lifetime, we can expect that the influence of L_1 on the process of forming hypotheses about L_2 will vary a great deal not only from individual to individual but also within a given individual.

Empirical evidence provided by second language acquisition research fully confirms that this is the case (Bertkau 1974, Taylor 1975, Schwarte 1982) and that the most obvious type of L_1 influence, ie. interlingual interference leading to interference errors, varies to a very large extent across individuals and across different acquisitional stages within the same individual. Also well known studies by Kellerman (1978) and Jordens (1978) confirm that a given learner's individual perception of relative distance between L_1 and L_2 is an important factor determining the extent of transfer from the native language.

Again, to be quite fair towards the Contrastive Analysis Hypothesis we have to note that it was not incompatible with the idea of variability across individuals as far as the effects of interlingual transfer were concerned, but this variability was seen only as the function of a number of successful responses and reinforcements and not as dependent on any internal, psychological factors.

The view that L_1 influences the process of hypothesis formation and testing can lead to another important conclusion, namely, that direct interlingual interference is only one of the possible manifestations of this influence and that other manifestations may be less direct and, as such, not easily observable. For example, this indirect influence may be manifested in the learner's tendency to avoid a given target language item in his utterances as much as possible, as was the case in the well-known Schachter's (1974) study, where the difficulty residing in the use of relative clauses in English as a second language was shown to be not so much inherent but rather related to the cognitive grid of a given L_1.

Even more importantly, this indirect influence can be discovered as a factor determining relative rapidity with which a given learner arrives at a given interim hypothesis and relative persistency with which he holds onto this hypothesis. This was very well demonstrated in Schumann's (1979) study. Schumann, looking at the appearance of pre-verb negation in the interlanguage of some learners of English representing various language backgrounds, found that the position of the negative in the first language did have influence on how long this stage persisted in a given student's interlanguage. For example, Spanish speakers used no + V extensively and persistently, while for students from L_1 with post-verb negation (e.g. Japanese) the no + V structure was of a very short duration.

Similar findings were reported in Huebner's (1979) case study of the acquisition of the English definite article by a Laotian learner.

Thus, as we can see, the adoption of the view that L_1 influences the process of hypothesis formation has more explanatory power than the Contrastive Analysis Hypothesis and can better account for second language acquisition data.

Let us now pass to the consideration of some pedagogical consequences of the view that second language acquisition is basically rule learning and let us compare these consequences with the pedagogical precepts derivable from the Contrastive Analysis Hypothesis and from the whole behavioristic concept of language acquisition. These considerations will refer to three important teaching strategies, namely, the strategies of error prevention, error correction, and error eradication.

As far as prevention of error is concerned, the Contrastive Analysis Hypothesis prompted two solutions. One of them consisted in focusing the teacher's attention on the areas of greatest structural divergence between L_1 and L_2 as revealed by contrastive analysis. Knowing these areas in advance, the teacher could provide a particularly great number of intensive habit formation techniques, mostly in the form of pattern practice drills, for the successful acquisition of these troublesome items.

The other solution, inherent in the whole behavioristic approach to language learning, consisted in keeping a strict control over the learner's productive output in the beginning stages of second language acquisition so as actually not to give the learner any chance of making mistakes.

Now, if we hold the view that hypothesis formation is what matters in second language acquisition, our pedagogical efforts will go not so much in the direction of inducing the learner to perform many correct repetitions of a given second language item but rather in the direction of maximally facilitating for him the process of hypothesizing about that item. To achieve this facilitation we should, first of all, try to predict what erroneous hypotheses, caused, for example, by overgeneralization or L_1 influence, the learner may set up. Then we should try not so much to prevent him from making such hypotheses (although this is to some extent possible by a well-planned presentation of the new item) but rather to provide him with plenty of opportunities for testing and correcting these hypotheses. For this purpose we will also construct exercises, although certainly not in the form of mechanical drills but rather in the form of problem-solving tasks which may even provoke the learner into making certain predictable types of mistakes. The belief in the effectiveness of this procedure stems from our acquisitional paradigm, according to which the learner profits greatly from making mistakes provided that he gets clear negative feedback.

As far as keeping control over the learner's productive output is concerned, it seems that the behavioristic notion of reinforcement and the behavioristic concern with consequences of the learner's verbal behavior have not necessarily been completely contradicted by second language acquisition research. Again, it seems that the behavioristic schema was a certain theoretical overgeneralization with reference to the present-day notion of hypothesis testing through feedback. It was an over-generalization in the sense that it emphasized the importance of external reinforcing stimuli originating in the person of an external agent. The cognitive notion of feedback is more subtle because, without denying the importance of positive or negative signals sent by the interlocutor, it also takes into account a host of personal factors residing in the learner, such as motivation, level of aspiration, subjectively accepted goal of language study, preferred learning style, etc. It seems that these internal factors often determine the level of performance accuracy reached by a given learner to a larger extent than giving or withdrawing reinforcement on the part of an external manipulator.

Some recent studies of feedback and the distinction between cognitive and affective feedback (Vigil and Oller 1976) must also lead to the conclusion that feedback is often ambiguous to the learner and is either misconstrued or ignored. This misconstruing or ignoring of feedback causes fossilization. When we consider conditions under which feedback is often misconstrued or ignored, we may conclude that there are two such basic conditions which may appear jointly or separately. One of these conditions appears when the communicative function of language becomes the dominant goal of language acquisition and the social (identity marking) function is not included as one of worthwile goals. Particularly Schumann's (1978) studies on the pidginization of interlanguage are illustrative of this condition.

The other condition obtains when the learner is forced to use communication strategies (for the notion of communication strategy, see Tarone 1981), ie. when he has to outperform his actual competence. Under such circumstances it is most often impossible for the interlocutor to provide clear cognitive feedback referring to the correctness of language forms.

So, after all, it seems that the behavioristic fear of early uncontrolled output on the part of the learner can also be theoretically justified within the contemporary schema of hypothesis formation and testing. Accordingly, it is clear that when we want our learners to achieve a relatively high degree of correctness, particularly in the context of foreign language teaching, we must not allow them to practice communication strategies in the classroom. Our teaching strategy in this case should rely on gradual building up of the learner's competence based on mostly reproductive activities. In other words, we should always make sure first that our learner is in possession of linguistic means and of a good performance model before assigning to him a communicative task requiring spontaneous production.

Another pedagogical consequence of viewing the process of second language acquisition as hypothesis formation and testing is related to the strategy of correcting language errors in the classroom. The behavioristic strategy based on the stimulus-response-reinforcement schema consisted in the immediate providing of the correct form and in making the learner repeat it. This strategy is totally incompatible with the notion of rule learning since we understand today that the cognitive processes of forming hypotheses about the target language and correcting them are more essential to language acquisition than even a large number of mechanical repetitions of the correct form. Accordingly, the strategy of correcting errors stemming from the present-day acquisitional schema logically leads

to emphasizing the importance of the learner's self-correction. How this self-correction is induced depends primarily on the strategy of teaching grammar, ie. on the fact whether the teacher has adopted the explicit or the implicit approach. Working with the explicit approach, which relies on explicitly formulated pedagogical grammar rules, the teacher should first signal to the learner that an error has been committed and wait for him to self-correct. If this were not sufficient, the teacher should prompt by indicating the grammatical category that has been violated (saying, for instance, "tense" or "article"). If this still does not work, the teacher should explicitly refer the student to a given grammar rule. As far as the implicit approach is concerned, the method of error correction employed by the Silent Way is a good illustration. This method also involves signalling to the student that an error has been made and waiting for him to self-correct; if the student is not able to self-correct, peer correction is supposed to take place.

Finally, let us consider the third pedagogical consequence of our acquisitional paradigm related to the problem of eradicating error. Let us note first that the behavioristic withdrawal of reinforcement, which in theory should have been quite sufficient for the successful extinction of an unwanted habit, did not work too well in the classroom and that is why the behaviorists came up with the idea of remedial exercises. Exercises of this sort involved many mechanical, correct repetitions of the problem item but it soon became evident that such repetitions were seldom sufficient for the abandoning of the old habit and the developing of a new one. And, again, the adoption of the view that language acquisition is rule learning offers a better solution, which seems to agree more with accumulated teaching experience. This solution consists in explicitly focusing the learner's attention on the erroneous item, in developing his motivation to acquire the correct form, and in providing him with a lot of meaningful practice involving the use of the item in question. Since the human mind has only a limited capacity for dealing with strictly controlled verbal activities, we can expect that the solution mentioned above is effective only when task overload is avoided, ie. when the learner can fairly easily focus on the problem item in the context of not very demanding production tasks.

To conclude, it can be stated that the contemporary concept of language acquisition, which views second language acquisition as the process of rule learning, can better account for acquisition data, especially in what refers to the influence of L_1, than the Contrastive Analysis Hypothesis. At the same time this concept prompts some pedagogical solutions referring to the strategies of error prevention, error correction and error eradication which are more in agreement with accumulated teaching experience and with some successful contemporary methods of language teaching than behavioristic recommendations.

REFERENCES

Bertkau, J. 1974. "Comprehension and production of English relative clauses in adult second language and child first language acquisition", Language Learning 24, 279-286.

Huebner, T. 1979. "Order of acquisition vs. dynamic paradigm: a comparison of methods in interlanguage research", TESOL Quarterly 13, 21-28.

Jordens, P., and E. Kellerman. 1978. "Investigation into the strategy of transfer in second language learning". Unpublished paper presented at AILA Congress, Hamburg, Germany.

Kellerman, E. 1978. "Give learners a break: native language intuitions as a source of predictions about transferability", Working Papers on Bilingualism 15, 59-92.

Marton, W. 1983. "Second language acquisition tactics and language pedagogy", System 11, 313-323.

Oller, J.W., and S.M. Ziahosseiny. 1970. "The contrastive analysis hypothesis and spelling errors", Language Learning 20, 183-189.

Schachter, J. 1974. "An error in error analysis", Language Learning 24, 205-214.

Schumann, J.H. 1978. The pidginization process: a model for second language acquisition. Rowley, Mass.: Newbury House.

Schumann, J. 1979. "The acquisition of English negation by speakers of Spanish: a review of the literature", in R. Andersen (ed.), The acquisition and use of Spanish and English as first and second languages, Washington, D.C.: TESOL, 3-32.

Schwarte, B.S. 1982. The acquisition of English sentential complementation by adult speakers of Finnish. Jyväskylä Cross-Language Studies 8. Jyväskylä: University of Jyväskylä.

Tarone, E. 1981. "Some thoughts on the notion of communication strategy", TESOL Quarterly 15, 285-295.

Taylor, B. 1975. "The use of ove-generalization and transfer learning strategies by elementary and intermediate students in ESL", Language Learning 25, 73-107.

Vigil, N.A., and J.W. Oller. 1976. "Rule fossilization: a tentative model", Language Learning 26, 281-295.

Wardhaugh, R. 1970. "The contrastive analysis hypothesis", TESOL Quarterly 4, 123-130.

Robert N. Vanderplank
Helsinki University Language Centre/Language
Centre for Finnish Universities, Jyväskylä

SEVEN PROBLEMS OF EVALUATION IN A UNIVERSITY LANGUAGE CENTRE[1]

Introduction

The purpose of this paper is to try and raise evaluation to a status which is higher than the one it holds at present. At the same time, I want to avoid treating the subject in too abstract or theoretical a manner and hope to bring it down to earth by discussing certain aspects of evaluation in terms of specific problems which I have noted during the past year at Helsinki University Language Centre.

The seven problems

Problem One: The 'testing vs. evaluation' problem

When I was asked if I would like to give a workshop session, I gave the request some thought and because I was reading a rather stimulating article on 'Evaluation', I jotted down this as the title of my own workshop. When the workshop timetable appeared, a colleague asked me about the title. It was, she declared, no more than a currently popular expression for testing - at best the theory behind testing. Being caught off guard, and certainly not having thought much about the content of my workshop at that point, I was naturally stung into action (or rather, into some hard thinking). Were they the same thing? Had I been playing word games? The more I thought, the less I agreed with my colleague. I had meant a much broader idea than the one we think of as 'testing'. I had meant the whole process of setting, judging, correcting, revising, up-dating by teachers and students, by administrators and colleagues. The list of tasks which I would bring under the umbrella of

---

[1] This paper is a revised version of one given at the English Oral Skills Workshop, University of Tampere, April 6 - 8, 1984.

evaluation grew over one weekend into the following list: (no particular order)

1. Giving students follow-up quizzes to check retention.
2. Asking students to state how much they have understood.
3. Asking students about how much they thought they had learnt.
4. Asking students to state/write their needs and objectives.
5. Asking colleagues what they think of one's tests, materials, objectives, etc.
6. Deciding what to do with poor materials and tests and when to update materials.
7. Deciding how, why, when and where students should be tested.
8. Deciding how much progress students should make.
9. Deciding what should be tested.
10. Deciding whether students have made any progress.
11. Deciding why or whether tests are/are not working.
12. Deciding whether techniques and assumptions are valid for a particular group.
13. Finding out whether students need English.
14. Finding out whether students have reached a criterion level.
15. Checking hunches on language learning.
16. Having what one does in class/language laboratory judged according to specific criteria.
17. Finding out whether students have learnt any English/made progress in English.
18. Having colleagues/outsiders observe and comment on your classes/ students.
19. Having students fill in questionnaires on their learning and your teaching - relevance, content, methods, etc.
20. Having students do something wrongly/badly - then seeing what went wrong, how to put to right and trying out one's hypothesis.
21. Selecting materials.
22. Grading materials.
23. Deciding whether one, as a language learner, has made any progress.

No doubt any experienced teacher could add a few more. What I think I have shown by this list is that testing is only a small part (shall we say the most formal part) of the whole spectrum of evaluation. If we take points 13, 14 and 17 as covering the formal placement, proficiency, continual assessment and achievement tests, then it is clear that testing holds its honored place because it is the point where the formal system of (University) administration takes an interest in the student. The administration wants to know certain things and it wants reliable information. But just because a body outside the classroom wants some information, this does not mean that all other aspects of evaluation are less important or less relevant - they may not be important to the administrators, but to my mind they are a central part of the teaching-learning process. In fact, I would go so far as to say the following: if motivation is a matter of perceived progress, and perceived progress is a matter of evaluating one's progress, then motivation and evaluation are rather closely linked.

If you are a language centre teacher, I think you will agree that the twenty-three points also represent many of the problems that we face in language centre teaching. Perhaps solutions to many of these day-to-day problems lie in raising teachers' and students' consciousness of evaluation and improving both the skills and knowledge involved in efficient and well-founded evaluation procedures. I hope that the following six problems will illustrate what I mean more fully.

Problem Two: The 'midnight oil' problem

Speaker A: Sorry, but these distractors are no good.
Speaker B: Oh §§§§§§§§§§§§§§ - but I was up until midnight doing these.

I think this is a frequent and serious problem. Teachers are asked to perform tasks for which they often have no training or an inadequate training. The quality of the work produced is only too easily equated with the quantity of time spent on it. In this particular case, the writer had spent hours constructing multiple choice items and resented being told that there was no point in giving them to the students in their present state since either they would not distract, or there were ambiguities, or they were simple yes/no items put together as a multiple choice question. There appears to be a contradiction in most language centres between the admitted lack of expertise on the part of many teachers and their willingness to involve themselves in tasks for which they are less than fully prepared.

The argument is often put forward that teachers have learnt a lot through spending hours developing materials and tests over the past few years in language centres, and that burning the midnight oil has meant that materials and tests have improved. I do not think that this argument holds water. At least not as far as the Language Centre and its clients, the students, are concerned. Personal development may be an admirable goal, and spending hours on writing and re-writing materials may be a laudable means of achieving that goal, but it is neither efficient nor in the best interests of the Language Centre and its students. If you know how to write multiple choice tests, you can avoid the 'midnight oil' problem - and, of course, spend time on more useful and interesting work. You could even spend time on keeping up-to-date with developments in testing and evaluation.

Problem Three: The 'something wrong' problem

Speaker C: There's something wrong with this listening comprehension test. I don't think it tests listening comprehension.

Speaker D: Sure it does, they have to listen to get the right answer.

Speaker: C: Yes, well, there's something wrong with it - but I'm not quite sure what it is.

Probably an even more common problem than Problem Two. After all, how many (language centre) teachers have the formal language and training to express their intuitive feelings of 'wrongness' in tests, materials and methodology. It is bad enough not to be able to tell a colleague something is wrong for fear of giving offence - but to know that something is wrong and to lack the means and knowledge for expressing what one knows is perhaps even more frustrating. A recent article by D.S. Taylor in BALT Journal[2] on the uses and value of Applied Linguistics makes precisely the same point. The formal disciplines of linguistics and its associated branches provide us with a reference framework and a language for illuminating and discussing, and perhaps even resolving, our constant problems of evaluation. I shall finish off this particular problem with another dialogue, which will illustrate another side of the same problem.

Speaker E: Shall we use this test again?

Speaker F: Well, didn't so-and-so say that he didn't like them very much.

Speaker E: Oh, what did he say exactly?

Speaker F: Well, I can't remember - but he said something about it being a general test.

---

[2] 'Linguistics, Applied Linguistics, and Language Teachers' in *The British Journal of Language Teaching*, Vol. 21, No. 3, 1983.

Problem Four: The 'so what?' problem

This problem is a development of Problem Three. Let us imagine that Speaker C finds out what was wrong with the test and tells Speaker D that the test is not really testing listening comprehension but is, for reasons X and Y, testing reading and recognition. Speaker D then replies 'So what? We've used this test as a listening comprehension test before and it worked well.'

There is, of course, not much that you can do unless you have some power and influence over Speaker D and can tell him/her to go away and do some homework. You can appeal to a sense of professionalism, to the critical judgement of others inside and outside, to a sense of honesty towards the students - but at the end of the day, there is not much you can do apart from offering to write another test which you think will really test listening comprehension.

I've come across so many 'so what?' problems of various types, but I shall only mention a few of them here. For example, there is the:

Speaker G: Look, this test doesn't discriminate between students at all well - they've all got between 19 and 24 out of 25.

Speaker H: So what - it doesn't really matter - it'll just push them all up a bit.

or the:

Speaker J: My God! These answers are all over the place - I don't think that a lot of these items are at all reliable.

Speaker K: Oh, it doesn't matter - I don't think it'll make any difference to the final result.

I think that once again many of these 'so what?' problems could have been avoided by careful selecting, grading and evaluating of test items and of testing materials. We have an obligation to know and understand what we are doing and why we are doing it - and we should be able to explain and justify our decisions and actions in the appropriate language when called upon to do so.

The 'so what?' problem leads on to:

Problem Five: The 'Rolls Royce Bumper' problem

This problem is named after a paper which I gave some years ago called 'Using the language laboratory or Rolls Royce bumpers make very good bottle openers' (Mextesol, Acapulco, 1980). While a tool (any tool, even a Rolls Royce bumper) may be very good for doing odd jobs, it may not be the purpose for which the tool was intended, and indeed it may be that its common use represents an enormous waste of valuable resources. Such an example of the 'Rolls Royce bumper' problem is to be found in the language laboratories of Helsinki University Language Centre. Let us say that roughly three-quarters of students' time is spent entirely on listening practice in the language laboratory. Why, then, install expensive language laboratories when listening centres cost about a quarter of the price? The uses of the language laboratories would not stand up to any evaluation in terms of the basic criteria for language laboratory use. That is, that it provides:

faster
better
more enjoyable[3]

development of speaking and listening skills than any other comparable teaching aid (e.g. tape recorder) or no aid. I would not deny for a moment that you can use the language laboratory for listening practice, but I would strongly deny that you are thereby making good use of valuable and limited resources.

There is perhaps a reluctance to get to grips with the process and findings of evaluation - an inertia combined with a lack of informed opinion about what to evaluate and how to evaluate. If we have such valuable and useful (at least potentially useful) resources, why are we not evaluating how best we can use them, how current approaches can be combined with the advantages of language laboratory use to produce maximum potential, instead of accepting the language laboratory as a mere vehicle for input - the minimum potential.

Problem Six: The 'good tape' problem

Problems Six and Seven deal with the problems of material selection and grading. Let us start with Speaker L.

Speaker L: Hey - here's a good tape - 'Acid Rain' - 15 minutes, interview format - could be relevant to science students - at least some of them - and it's a topical subject.

I am fully aware of the problems of obtaining relevant material. I am also aware of the problems of converting raw materials into teaching material.

---

[3] These are the basic criteria for the evaluation of any tool.

I just wonder at times if everyone else is aware of the second problem and the procedures that have been developed over the past 10 - 12 years for selecting and adapting materials for ESP and EAP teaching.

Now it seems to me that our job is not only to provide our students with practice in listening but also to train them in listening skills. The above criteria - topic, length, content, format - are all relevant to evaluating practice materials (rather like selecting the programmes you want to listen to on the radio). But practice is surely just an element included in all behaviour which we cannot carry out confidently. While it must contain an implicit training element, it is essentially a repeating and self-improving process and, above all, a self-evaluative process: 'Let's see if I can do it better this time!'. Evaluating and adapting materials for training is a far more complex and demanding process than selecting materials for practice. Factors involving the accent, stress and rhythm patterns of the speakers, the communicative purpose of the material, the cohesion and coherence of the discussion/interview, the contribution of the material to the development of the learners' listening ability in terms of listening skills, the identification of specific problems and focus on them - all these factors and others need to be considered. Indeed, the whole process of adaptation may turn the original practice criteria on their head. That is, why is this a bad or difficult piece of listening? What problems will the students have with it? What training can we devise that will help them overcome these problems and also help them cope with such pieces of listening in future. Effective language training depends, to a large extent, upon systematic and constant evaluation of materials, methods and, above all, of the learners' progress: 'Are my training materials and methods helping the students to overcome their problems?'.

Problem Seven: The 'just what we need' problem

Problem Seven is closely related to Problem Six. I am thinking of the following situation:

'Great, just what we need, Professor Muggins talking about the sex life of pigeons in Trafalgar Square.'

Our students are very good at taking in large quantities of information, especially in their own subject areas. I was surprised, quite recently, to hear some Law students saying how much they had enjoyed hearing a recorded talk on the Civil Courts of England and Wales. There is clearly no accounting for taste - but then, after all, they are Law students. The point I wish to make is that the information in the talk was interesting, perhaps even useful to some at a later date, but there was no reason why it should have been recorded on tape. The recording itself had no particular value - it was written text read aloud. It contained very few of the features of normal spoken English, it did not even sound like a lecture (if there can be such a thing as a normal lecture). You really could not have analysed the talk as a piece of spoken discourse as it had so few of the features of spoken discourse. As far as I could judge, it had been selected primarily (perhaps only) for its topic and information content. If one wants to learn about the Civil Courts of England and Wales, there are perhaps better ways of doing it than sitting in the language laboratory for an hour. Maybe our students are not aware of that - maybe we should tell them.

This topic-content-practice vs. communicative purpose-language-listening training problem comes up again and again and I think we have to be very clear about the place of each and our role in providing each of them. While a Law specialist may think that listening to a talk on some aspect of Law is useful practice, with so little time available the language teaching specialist may well think that the students should concentrate not on the content of the talk - which may cause few problems, anyway - but on the presentation, the method by which the information is provided, the whole argument structure of the talk and the process through which our understanding is built up. The language teaching specialist may even decide that while the content is clear, the language of the talk is so obscure and unnecessarily complex (that is, the whole talk is so poorly presented) that in language learning terms there is little to be gained by using it in its actual form.

Last thoughts

I have presented, in a rather rambling fashion, some of the problems which I have come across during the past year at Helsinki University Language Centre. As I said earlier, I feel that many of these problems could have been avoided by systematic and solidly-based evaluation procedures. By that I mean that there is an admitted lack of expertise and training in many of

the areas discussed above, and, perhaps more important, no policy of recruiting staff who are already trained in what are clearly the disciplines of Applied Linguistics.

I said at the beginning of this paper that I wanted to raise the status of evaluation. In Finland, the best way of raising the status of anything in language learning and teaching is to call it 'research' and to call for more research in 'it'. I do not want to do that. I do not think that research necessarily solves or even helps with our day-to-day problems. I would like to suggest, however, that evaluation is the closest many of us will ever come to research and experimentation. I do not think that we should be going off into elaborate research projects in Language Centres just because it is the only way to make them important and raise their status, but I do think that there is a great deal of work on the evaluation of tests, materials and methods just lying around waiting to be turned into Licentiates and Doctorates - solid work of real practical value - but perhaps those interested should do a course in Applied Linguistics first - it will save a lot of time later.

Liisa Korpimies
Language Centre for Finnish Universities

ZUR EINSCHÄTZUNG DES FREMDSPRACHENUNTERRICHTS AN SPRACHENZENTREN AUS DER SICHT DER STUDENTEN

1. Vorwort

Die vorliegende Untersuchung ist Teil einer im Rahmen des Zentralen Spracheninstituts der finnischen Hochschulen durchgeführten Folgestudie zum Fremdsprachenunterricht [1]. Hauptziel des Forschungsprojektes ist es herauszufinden, wie gut die vorgegebenen Sprachfertigkeiten durch den Fremdsprachenunterricht an Sprachenzentren erreicht werden.

Die Untersuchung gliedert sich in mehrere Teile. Der erste, des Verbundsystems der Sprachenzentren, ist bereits abgeschlossen (Korpimies und Utriainen 1982). Das zweite Teilprojekt, das die Ziele des Sprachenunterrichts an Sprachenzentren und ihre Präzisierung zum Gegenstand hat, soll allgemeine und fachspezifische Ziele des Fremdsprachenunterrichts möglichst genau bestimmen. Das soll der Planung und Durchführung des Fremdsprachenunterrichts dienen sowie Kriterien für Bewertungsmaßstäbe liefern. Dieses Teilprojekt ist bereits angelaufen. Dazu und zum folgenden Teilprojekt gehört die jetzt fertiggestellte Einschätzung des Unterrichts an Sprachenzentren durch die Studenten, die also die Durchführung einer Evaluation zum Inhalt hat. Gegenstand des dritten Teilprojekts, der Bewertung des Standes von Sprachfertigkeiten, ist es, zu untersuchen, wie gut die präzisierten Ziele des Sprachenunterrichts an Sprachenzentren in der Zeit erreicht werden können, in der die Studenten and den Sprachkursen der Sprachenzentren teilnehmen können. Das vierte und letzte Teilprojekt hat die Bewertung des bleibenden Einflusses zum Inhalt. Ihr Ziel ist es einzuschätzen, ob die von Studenten erreichte Sprachfertigkeit erhalten geblieben ist und ob er sie während seines Studiums und im Arbeitsleben nutzen konnte. Die Evaluation setzt somit langfristige Untersuchungen voraus.

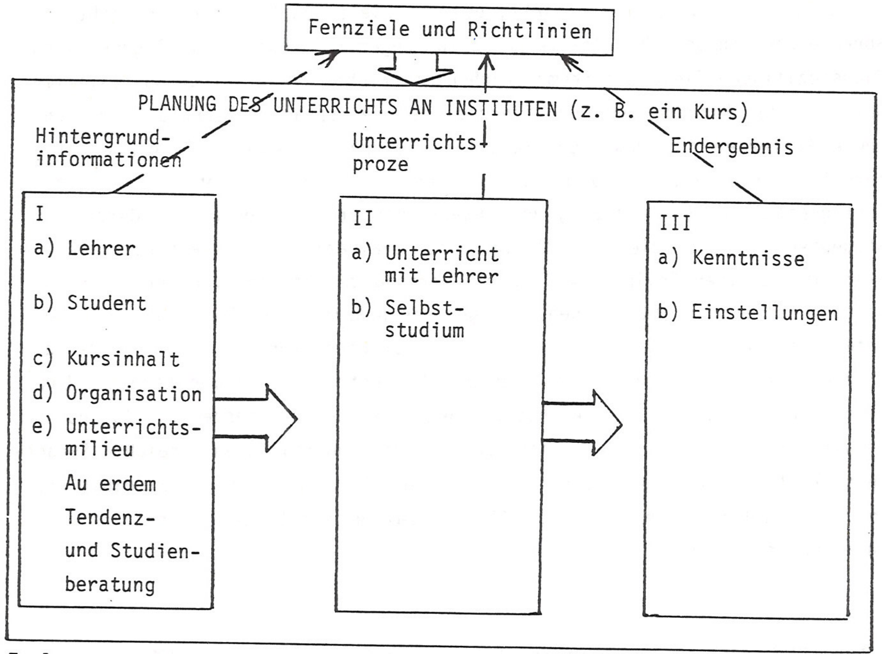
---

[1] Korpimies - Valtanen - Vaherva: Kielikeskusopetus opiskelijoiden arvioimana. Kielikeskusopetuksen seurantatutkimus, II vaihe.

2. Einleitung und Rahmen der Forschung

Evaluationsstudien werden im Schulbereich schon lange und in geringem Maße auch an Hochschulen durchgeführt. In der letzten zehn Jahren verschob sich der Schwerpunkt von der Bewertung des Endresultats eindeutig in Richtung auf Untersuchungen des Unterrichtsprozesses. Eine umfassende Evaluation muß alle Teilbereiche (Aktionsrahmen, Einsatz, Prozeß und Resultate) einschließen. In diesem Rahmen soll auch der Fremdsprachenunterricht an Sprachenzentren eingeschätzt werden.

Das Unterrichtsgeschehen kann als Prozeß der aktiven Wissensaneignung verstanden werden, in dem der Student nicht Objekt, sondern Subjekt ist. Deshalb muß dem Fremdsprachenunterricht und seiner Anwendbarkeit aus der Sicht der Studenten große Bedeutung bei der Auswahl der Evaluationsmethode beigemessen werden.



Evaluationsmodell nach Franke-Wikberg und Johansson (1976).

In unserem Folgestudieprojekt soll nach oben angegebenem Modell verfahren werden. In dem bereits fertiggestellten Teilbericht (Korpimies & Utriainen 1981) sind die unter 'Hintergrundinformationen' zusammengestellten Informationen beschrieben. Z. Z. wird die Tendenzbeschreibung des Fremdsprachenunterrichts an Sprachenzentren präzisiert, die implizit bereits Einfluß hatte auf den Unterricht und die Prozeßevaluation und die man bei der Erstellung der Bewertungsmaßstäbe (Sprachtests) und der abschließenden Resultatevaluation nutzt. Wie im allgemeinen bei solchen Evaluationsstudien konzentriert sich auch dieses Projekt nur auf einen Teil der Gesamtheit, wobei andere Komponenten der Gesamtheit mehr oder weniger berührt werden. Der vorliegende Report berichtet über die Ergebnisse der Prozeßevaluation, betrachtet aber auch andere Teil der Gesamtheit.

2.1. Zielstellungen und Herangehen an das Teilprojekt

In der zweiten Hälfte der 70er Jahre wurde der Fremdsprachenunterricht für Studenten aller Fachrichtungen im Rahmen der Hochschulreform erneuert. Seitdem wurde die Verantwortung des zum Studium gehörenden obligatorischen Fremdsprachenunterrichts neuen Einrichtungen, den Sprachenzentren übertragen. Die Umsetzung der Reform in die Praxis erfolgte allmählich, und die Fakultäten/Studienpläne paßten sich ihr ebenfalls allmählich an. Jetzt, nach beinahe zehn Jahren seit der Reform, stabilisiert sich der Fremdsprachenunterricht: fast alle Hochschulen haben ein eigenes Sprachenzentrum und fast alle Studenten studieren nach dem neuen System.

Da sich der neue Fremdsprachenunterricht nur schrittweise durchsetzte, war es schwierig, sich zu einem früheren Zeitpunkt ein umfassendes Bild über seine Effektivität zu machen. Jetzt, bevor sich der neue Unterricht letztendlich stabilisiert, wird es notwendig, über die Planung und Durchsetzung des neuen Fremdsprachenunterrichts Bilanz zu ziehen.

Die hier vorliegende Studie hat das Ziel, die Sprachkurse aus der Sicht der Studenten zu untersuchen. Obwohl finnische Studenten nur wenig an der Evaluierung des Unterrichts teilhaben, sind sie in der Lage, den Unterricht sehr gut zu beurteilen. Ziel der Untersuchung ist vor allem, ein Feedback von den Studenten für die Planung des Fremdsprachenunterrichts zu erhalten, aber auch Informationen für die eigene Entwicklung der Lehrer. Mit Hilfe der Studie soll herausgefunden werden, wie effektiv der Fremdsprachenunterricht sowohl während des Studiums als auch später genutzt werden kann. Als Hintergrund für die Meinungen werden frühere Sprachfertigkeiten und die neben dem

Sprachstudium gleichzeitig verlaufenden anderen Studien herangezogen. Gleichzeitig werden die Studenten gebeten, Wünsche und Verbesserungsvorschläge zu den Sprachkursen und Sprachstudien zu äußern.

Die Befragung der Studenten soll die wichtigsten Punkte im obligatorischen Fremdsprachenunterricht in der Phase aufklären, wo eine Gestaltung der Kurse nach der Wünschen der Studenten noch möglich ist.

3. Probleme

Problem 1. Wer studiert an Sprachenzentren?

Zuerst wird analysiert, wann die Studenten ihr Studium an der Hochschule begonnen, wieviel Jahre Sprachunterricht sie in der Schule in der von ihnen gewählten Sprache absolviert und welche Note sie im Reifezeugnis erhalten haben sowie ob sie die betreffende Sprache außerhalb der Schule bzw. beim Aufenthalt im Ausland anwenden konnten. Außerdem werden die Studenten um eine Einschätzung ihrer mündlichen und schriftlichen Sprachfertigkeiten zu Beginn des Sprachkurses gebeten.

Problem 2. Wie werden die Sprachkurse organisiert und wie ist die Einstellung der Studenten?

Daraufhin wird die praktische Durchführung der Sprachkurse untersucht: wie sind die Kurse zeitlich angelegt, ist dies eine für die Studenten günstige Zeit, wie aktiv ist die Teilnahme am Unterricht und was sind die Gründe für eventuelle Fluktuationen. Von Interesse ist außerdem, ob Hausaufgaben gefordert werden und wieviel Zeit sie in Anspruch nehmen. Die Studenten werden auch nach Eignung der Unterrichtsräume befragt, welche Hilfsmittel, Unterrichtsmaterialien und Arbeitsformen im Kurs angewendet werden und ob die Arbeitsweisen dem Ziel des Kurses entsprechen.

Problem 3. Wie effekktiv schätzen die Studenten den Sprachunterricht ein?

Teilproblem 1. Wie nützlich sind die im Sprachkurs erworbenen Fähigkeiten für das Studium, den zukünftigen Beruf und die Entwicklung der Sprachfähigkeit?

Der Nutzen der erworbenen Fähigkeiten wird mit Hilfe der folgenden Fragen untersucht: halten die Studenten die Sprachstudien für notwendig, haben sich ihre Sprachkenntnisse im Laufe des Kurses verbessert, kann das im Kurs Gelernte im Studium und zukünftigem Beruf verwendet werden und haben die Studenten ihrer Meinung nach neues gelernt.

Teilproblem 2. Wie motivierend und von der Atmosphäre her positiv empfinden die Studenten die Sprachkurse?

Die Motivation und dazugehörende Momente und Faktoren werden an der allgemeinen Zufriendenheit der Studenten mit dem Hochschulstudium gemessen sowie daran, wie begeisternd, interessant oder frustrierende das Studium im Sprachkurs erlebt wurde und ob er die Erwartungen der Studenten erfüllt hat. Weiterhin sollen die Studenten die Studienatmosphäre, die Aktivität ihrer Kommilitonen sowie das Verhältnis Lehrer-Studenten einschätzen.

Problem 4. Wie schätzen die Studenten den Unterricht am Sprachenzentrum ein?

Teilproblem 1. Wie wird die Zielsetzung im Unterricht angestrebt?

Hierher gehören Fragen wie: Wurden die Ziele und Bewertungsgrundlagen den Studenten bekanntgemacht, hätten sich die Studenten mehr Informationen über die Kursziele gewünscht oder sind diese Ziele und Bewertungsgrundlagen unklar geblieben?

Teilproblem 2. Wie schätzen die Studenten Unterrichtsorganisation und ihre Durchsetzung ein?

Unterrichtsorganisation und ihre Durchsetzung werden geprüft durch Meinungsumfragen unter den Studenten über Größe der Studiengruppe, Schwierigkeitsgrad des Unterrichtsmaterials und seine Zweckmäßigkeit hinsichtlich der Studienfächer der Studenten, Effektivität der Ubungen, Zeitaufteilung des Unterrichts sowie seine Klarheit und Praxisrelevanz. Die Bewertung der Fähigkeiten des Lehrers durch die Studenten wird anhand der Stundenvorbereitung, Geschicklichkeit des Lehrers zu motivieren, das Interesse aufrechtzuerhalten und sich auf Wesentliches zu konzentrieren sowie seines Fachwissens und seiner Sprachfertigkeit gemessen. Die studenten geben auch ein Gesamturteil sowohl über den Kurs als auch über den Lehrer ab.

Teilproblem 3. Kann man ein Feedback erhalten und wie können Fortschritte der Studenten festgestellt werden?

Wodurch erfolgt das Feedback, haben die Studenten ihrer Meinung nach genügend Informationen über ihre Fortschritte erhalten und hat der Lehrer das Feedback im Unterricht genutzt sind Fragen, die dieses Teilproblem zur Untersuchung heranzieht.

Teilproblem 4. Was haben die Studenten ihrer Meinung nach im Sprachkurs gelernt und wie wollen sie die erreichten Sprachfertigkeiten aufrechterhalten?

Die Bewertung des Gelernten durch die Studenten erfolgt durch Selbsteinschätzung ihrer Sprachfertigkeiten auf den verschiedenen Teilgebieten (Textverstehen, Wortschatz, Beherrschung der Struktur und Terminologie im eigenen Fachgebiet, Übersetzungsfähigkeit, schriftliche Leistungen, verstehendes Hören, Sprechen und Aussprache). Es werden Umfragen dazu, wie die Studenten die erreichten Sprachfähigkeiten aufrechterhalten wollen, durchgeführt.

Problem 5. Welche Unterrichtszüge werden bei der Kursevaluation betont und welche Unterschiede werden bei der Evaluation der kurstypmäßigen Gruppen gefunden?

Die den Unterricht betreffende Dimensionen der Evaluation werden mit Hilfe der Faktorenanalyse zum gesamten Material sowie den verschiedenen Kurstypen untersucht und Gruppenunterschiede mit Hilfe einer Varianzanalyse geklärt.

Problem 6. Welche Wünsche und Verbesserungsvorschläge haben die Studenten in Bezug auf Sprachkurse und Sprachstudium?

Die Wünsch und Verbesserungsvorschläge der Studenten werden daran gemessen, was ihrer Meinung nach während des Kurses zu sehr oder zu wenig betont worden ist, was ihnen ihrer Meinung nach am meisten gegeben hat und wo sie die Gründe für eventuelles Mißlingen sehen sowie welche Wünsche und Verbesserungsvorschläge sie haben.

4. Durchführung der Analyse

Zielgruppen der Untersuchung waren die Studenten der schwedischen, englischen und deutschen Sprachkurse der Universitäten in Tampere, Oulu, Jyväskylä und Joensuu und der Technischen Hochschule Tampere. Unter Sprachkurse werden solche Kurse verstanden, die Sprachstudien betreffen, die im Studienplan der Studenten vorgesehen sind. Zu den obligatorischen Sprachstudien gehören in allen Studienplänen Unterricht in der zweiten Landessprache sowie der Erwerb schriftlicher und mündlicher Fertigkeiten in einer weiteren Fremdsprache. Diese ist in den meisten fällen Englisch, seltener Deutsch oder andere Sprachen. Einige Studienpläne verlangen auch Fertigkeiten (meist schriftliche) in einer zweiten Fremdsprache und das ist in den meisten Fällen

Deutsch. Die Auswahl der schwedischen, englischen und deutschen Sprache als Forschungsobjekt gründet darin, daß für ihren Unterricht die größten Reserven eingesetzt werden (Korpimies & Utriainen, 1981).

Folgende Faktoren bestimmten die Auswahl der untersuchten Kurse dieser drei Sprachen:

1.) Erhebungen wurden z. B. verschiedene Kurstypen betreffend gemacht (Textverstehen, mündliche Fähigkeit, integrierter Unterricht);
2.) Für die Erhebungen sollten Studenten möglichst vieler Studienrichtungen herangezogen werden; und
3.) Zusammenstellung des Material gemäß vorher vereinbarter Termine, was zum Teil die Verwirklichung der beiden oben genannten Punkte einschränkte.

Wie Tabelle 1 zu entnehmen ist, erhielten wir von Kursteilnehmern der schwedischen und englischen Sprache annähernd gleichviel Antworten, der Teil der Studenten der deutschen Sprache blieb geringer, da an den gewählten Prüfungstagen weniger Deutschunterricht gewählt worden war. Insgesamt erhielten wir 766 Antworten aus 68 verschiedenen Sprachkursen.

Tabelle 1. Aufteilung der Antworten nach Sprachen und Kurstypen

	schriftliche Fertigkeit		mündliche Fertigkeit		integrierter Unterricht		insgesamt
	Kurse	N	Kurse	N	Kurse	N	insgesamt
Schwedisch	10	136	4	26	14	195	357
Englisch	13	113	12	127	7	84	324
Deutsch	4	30	2	27	2	28	85
insgesamt	27	279	18	180	23	307	766

Schwankungen bei den Kurstypen nach Sprachen rühren daher, daß der Schwedischunterricht in Sprachenzentren häufiger intergriert ist als der Englisch- und Deutschunterricht, bei denen Unterricht zur Entwicklung des Textverstehens und mündlicher Unterricht gewöhnlich getrennt sind.

Genaue Angaben zur Zahl der Studenten an Sprachenzentren konnten nur sehr schwer festgestellt werden. In der von Korpimies und Utriainen (s.o.) durchgeführten Analyse wurden die Sprachenzentren gebeten, die Zahl ihrer Studenten zu schätzen, aber nur wenige konnten überhaupt ungefähre Angaben zur Gesamtzahl machen und Angaben zu Studenten verschiedener Sprachen fehlten ganz. Deshalb ist es schwer zu sagen, welchen Prozentsatz die von uns befragten Studenten an der Gesamtzahl aller Schwedisch-, Englisch- und Deutschstudenten ausmachten.

Die Erhebung ist auch dahingehend unvollständig, als daß sie nicht die Studenten aller Sprachenzentren erfaßt. Deshalb sind Studenten einiger Studienrichtungen entweder total außerhalb der Studie geblieben (z. B. Studenten der Rechtswissenschaften und der Land- und Forstwirtschaft) oder ihr Anteil ist sehr gering (z. B. Studenten der Medizin und des Lehrerbildungsinstituts). Da unsere Untersuchung nicht das Ziel hat, den Unterricht an verschiedenen Sprachenzentren an sich oder einzelne Sprachkurse zu analysieren, sondern den Unterricht aus studentischer Sicht zu betrachten und ihren Gesamteindruck dazu zu untersuchen, hat die Unvollständigkeit der Erhebung in diesem Fall keinen störenden Einfluß auf die Forschungproblematik.

Das verwendete Auswahlverfahren beruht aus praktischen Gründen auf Ermessen. Feldman (1978) bemerkt dazu, daß es immer problematisch ist, zuverlässige Daten in solchen Meinungsforschungen zu erhalten, da die Studenten nicht zufällig auf die Kurse verteilt sind; es ist außerdem schwierig, gerade die Population zu bestimmen, in der die Studenten des Kurses tatsächlich eine zufällige Auswahl bilden. In unsere Erhebung z. B. nehmen die Studenten des ersten studienjahres 36 %, die des zweiten 20 % und des dritten 24 % ein, die restlichen 20 % haben ihr Studium vor 1980 begonnen. Als Population wird in derartigen Untersuchungen oft eine undefinierbare Studentenpopulation betrachtet, die "beobachtungsartig" ist ("like those observed") (s.o.).

Eine in Form einer Studentenumfrage gemache Kursanalyse sollte nach Berkn (1979) rund fünfmal soviel Studenten betragen, wie der Meinungsumfragebogen Fragen enthält; dieses Verhältnis garantiert die Zuverlässigkiet der Ergebnisse.

5. Zum Umfang der Evaluation

Zur Einschätzung des Umfangs der Evaluation wurden 55 - 108 Varianzen in einer Faktorenanalye mit 5, 6 bzw. 7 Faktoren eingeschätzt. Die 6-Faktoren-Variante erwies sich für die Interpretation als am günstigsten und beinhaltete 44 % aller Varianzen. Die verënderlichen Kommunalitäten schwankten zwischen .03 - .65.

Folgende Faktoren wurden untersucht:
1. Faktor des Wohlbefindens
2. Faktor der didaktischen Fähigkeiten
3. Motivationsfaktor
4. Erfolgsfaktor

5. Zielfaktor
6. Faktor der Allgemeinen Zufriendenheit

Bei der Unterschuchung der verschiedenen Kurstypen mit Hilfe der Faktorenanalyse wurden dieselben Faktoren 1. - 5. eingesetzt, der 6. Faktor beschrieb die Unterrichtsorganisation.

5.1. Wer studiert an Sprachenzentren?

Zusammenfassend kann gesagt werden, daß der "typische" Student an Sprachenzentren meist im 1. oder 2., seltener im 3. oder einem höheren Studienjahr studiert. Englischstudenten haben gewöhnlich 10 Jahre, Schwedischstudenten 6 oder 10 und Deutschstudenten 3, seltener 6 Jahre die entsprechende Fremdsprache in der Schule gelernt. Sie haben im allgemeinen die Reifeprüfung in dieser Sprache abgelegt und über 60 % erhielten dafür die Prädikate magna cum laude oder laudatur. Nur selten haben sie sich über den Schulunterricht hinausgehende Sprachfertigkeiten erworben und nur 10 % von ihnen haben sich längere Zeit im Zielsprachenland aufgehalten. Zusätzlich zum Fremdsprachenunterricht haben die Studenten durchschnittlich 21 - 30 oder über 30 Stunden Fachunterricht pro Woche. Ihre eigenen Sprachfertigkeiten, besonders die mündlichen, schätzen die Studenten an Sprachenzentren zu Beginn des Sprachkurses recht bescheiden ein.

5.2 Sind Unterricht und Studium zweckentsprechend?

Ob Unterricht und Studium zweckentsprechend sind, wurde in vorliegender Untersuchung von zwei Gesichtspunkten aus betrachtet: von der Anwendbarkeit des Gelernten und von der Fähigkeit des Unterrichts zu motivieren. Die Antworten ergeben, daß die Mehrzahl der Studenten die Sprachkurse als notwendig und die in ihnen erworbenen Fähigkeiten als nützlich sowohl für den zukünftigen Beruf als auch für das Studium erachtet. Für das Studium wurde Englisch als nützlichste Sprache eingeschätzt, während alle drei Sprachen als gleich wichtig für den zukünftigen Beruf gehalten werden. Über die Hälfte der Befragten gab an, daß die Sprachkurse zur Verbesserung ihrer Sprachkenntnisse beigetragen haben. Die integrierten Kurse der verschiedenen Sprachen waren in dieser Beziehung am effektivsten. Über 80 % der Befragten hatten ihrer Meinung nach neues im Kurs gelernt und auch in dieser Beziehung waren die integrierten Kurse wirksamer als die anderen.

Die Motivation der Studenten wurde nach drei Geschictspunkten untersucht: nach der allgemeinen Zufriedenheit der Studenten mit ihrem Studium an der Hochschule, danach, ob der Sprachkurs das Sprachstudium motivierend beeinflußt und nach den das Wohlbefinden im Kurs bestimmenden Faktoren. Die Antworten ergeben, daß die Studenten der Sprachenzentren im allgemeinen recht zufrieden mit dem von ihnen gewählten Studienprogramm sind und das Studium an der Hochschule als sinnvoll betrachten; mit dem bisher erreichten Studienerfolg sind sie trotzdem nicht ganz zufrieden.

Die Motivationsfähigkeit der Sprachkurse wurde untersucht, indem die Studenten angeben sollten, ob sie auch an einem fakultativen Kurs teilnehmen würden. Über die Hälfte der Befragten gab an, wahrscheinlich teilzunehmen und die meisten Freiwilligen waren Teilnehmer integrierter Kurse. Die meisten ablehnenden Antworten kamen von Teilnehmern schriftlicher Kurse. Etwa ein Fünftel der Studenten empfand den Kurs als unnütz, wobei auch hier verhältnismäßig mehr negative Antworten aus schriftlichen Kursen kamen. Ungefähr die Hälfte der Befragten war bereit, ihre Sprachstudien auch in der Zukunft in irgendeiner Form fortzusetzen; am interessiertesten daran waren die Studenten der deutschen Sprache. Jeder Dritte war der Meinung, daß der Kurs dazu beigetragen hat, das Interesse an einer weiteren Vervollkommung seiner Sprachkenntnisse zu wecken und nur 5 % der Studenten gab an, daß sich ihr Interesse am Fremdsprachenunterricht auf den Kurs beschränkte. Etwa die Hälfte der Befragten war der Meinung, daß das im Kurs Gelernte ihren Erwartungen entsprach.

Der größte Teil der Studenten hatte den Sprachkurs als positiv empfunden. Nur knapp jeder fünfte hielt die Unterrichtsstunden für monoton und langweilig; etwas größer war der Anteil jener, die angaben, den Unterricht nur widerwillig zu besuchen. Schriftliche Kurse wurden von der Atmosphäre her weniger angenehm und ihre Gruppen passiver als andere eingeschätzt. Die Schwedischstudenten empfanden den Kurs und die Unterrichtsstunden etwas negativer als andere. Das Verhältnis Lehrer - Sudenten wurde für alle Sprachen freundschaftlich und warm eingeschätzt; zwei Drittel der Befragten beurteilten den Lehrer ermutigend und anspornend. Die Studenten konnten ihrer Meinung nach auch genügend Fragen während des Unterrichts stellen.

Verallgemeinernd betrachtet nahmen die Studenten eine sehr positive Haltung zu den Sprachkursen ein, einige Unterschiede fanden sich bei der Untersuchung sowohl nach Sprachen als auch nach Kurstypen. Die Unterschiede zwischen den Sprachen waren nicht groß, wenn auch der Schwedischunterricht nicht so begeistern konnte wie Englisch- oder Deutschunterricht. Schriftliche Kurse wurden im allgemeinen schwächer bewertet: sie wurden öfter als fruchtlos und eintönig empfunden. Integrierte Kurse dagegen wurden als angenehmer, nützlicher und begeisternder beschrieben.

5.3. Evaluation des Unterrichts an Sprachenzentren durch die Studenten

Die Evaluation des Sprachenunterrichts an Sprachenzentren wurde nach fünf Sachbereichen behandelt: Ziel des Unterrichts, Organisation und Durchsetzung des Unterrichts, Folgestudie zum Feedback und den Fortschritten, Evaluation der eigenen Fortschritte und der Aufrechterhaltung der Sprachfertigkeiten nach dem Kurs durch die Studenten im Fremdsprachenunterricht seine Meinungen zu den besten Auswirkungen des Kurses und/oder den Grünen für ihr Mißlingen.

Die Zielstellungen des Unterrichts wurden analysiert mit Hilfe der Antworten der Studenten zu Zielen und Evaluationskriterien der Kurse. Drei Viertel aller Schwedisch- und Englisch- und ca. die Hälfte der Deutschstudenten gab an, daß der Lehrer die Ziele in des Unterrichts erklärt hatte. Schwächer waren die Ziele in mündlichen Kursen erklärt worden. Jeder Fünfte Befragte wünschte sich mehr Wissen über die Ziele des Kurses. Die Evaluationsgrundlagen hatte der Lehrer der Hälfte der Befragten auseinandergesetzt; am schwächsten war die Situation im Deutschunterricht und in mündlichen Kursen. 40 % der Studenten waren die Evaluationskriterien unklar geblieben.

Mit der Organisation und Durchsetzung der Sprachkurse waren die Befragten recht zufrieden. Eine Verschiebung des Kurses auf eine spätere Phase im Studium wurde nicht gewünscht und die Größe der Studiengruppe wurde als geeignet eingeschätzt. Die Studenten der mündlichen Kurse konnten mehr als andere die Wahl der Unterrichtsmaterialien beeinflussen und der Schwierigkeitsgrad der Materalien wurde im allgemeinen als angemessenen betrachtet. Die Unterrichtsmaterialien waren dem Fachstudium ziemlich gut angepaßt; am besten wurde dies im Schwedischunterricht erreicht. Auch die Fähigkeiten der Lehrer wurden hoch eingeschätzt. Über 70 % der Befragten war der Meinung, daß der Lehrer gut vorbereitet war. Es war den Lehrern gelungen, das Interesse der Studenten am Unterricht recht gut aufrechtzuerhalten und im allgemeinen wurde nicht zu schnell fortgeschritten. Die Sprachfertigkeiten der Lehrer wurden durchweg als gut eingeschätzt und es wurde ihnen das Beherrschen des an die Unterrichtsmaterialien gebundenen Fachwissens bestätigt.

Die Befragten wurden außerdem gebeten, die Durchsetzung des Unterrichts und die Lehrer einzuschätzen. Die Bewertung fiel für alle Sprachen gleichmäßig positiv aus. Schriftliche Kurse wurden etwas weniger gut als die anderen Kurstypen bewertet. Die Studenten gaben auch eine Gesamteinschätzung des Kurses und des Lehrers ab: dabei erhielten die Lehrer eindeutig höhere Werte als die Kurse. Integrierte Kurse wurden besser als andere bewertet und die Lehrer mündlicher Kurse höher als die anderen eingeschätzt.

Die zum Erhalten von Feedback und zum Beobachten der Fortschritte eingesetzten Mittel unterschieden sich etwas nach Sprachen und Kurstypen, was sich jedoch aus den unterschiedlichen inhaltlichen Schwerpunkten der Kurse erklären läßt. 12 % der Befragten gab an, daß sie kein Feedback zu ihren Fortschritten erhalten hatten und fast die Hälfte war der Meinung, daß sie nicht genügend Informationen über ihre Fortschritte erhalten hatten; nur knapp 10 % der Befragten hielt das vom Lehrer erhaltenen Feedback für ausreichend. Die Mehrheit der Studenten war nach wie vor der Meinung, daß der Lehrer die zu den Fortschritten erhaltenen Informationen nicht nutzte, z. B. für die Differenzierung des Unterrichts.

Es ergab sich, daß die Einschätzungen der Studenten zur Entwicklung der einzelnen Sprachfertigkeiten (Textverstehen, Wortschatz, Beherrschen von Strukturen und Terminologie, Übersetzen, Schreiben, Verstehendes Hören, Aussprache und Sprechen) schwer zu interpretieren waren. Generell waren jedoch die zentralen Fertigkeiten, wie mündliches und schriftliches Textverstehen sowie die Beherrschung der Terminologie des eigenen Fachgebietes am besten entwickelt. Die Einschätzungen der Studenten waren recht vorsichtig: etwa die Hälfte der Antworten lautete "in gewissem Maße". Das Aufrechterhalten der Sprachfertigkeiten nach dem Kurs erfolgte hauptsächlich durch Literaturstudium, besonders durch das Lesen von Lehrbüchern. Nur wenige glaubten, auch ihre mündlichen Sprachfertigkeiten nach dem Kurs aufrechterhalten zu können.

Die Studenten schriftlicher und integrierter Kurse schätzten das Anhäufen des Wortschatzes des eigenen Fachgebietes und die allgemeine Entwicklung ihrer Sprachfertigkeiten am wertvollsten ein. Auch die mündlichen Schwedischkurse wurden lobend erwähnt für das Aneignen der Terminologie des eigenen Fachgebietes. Den mündlichen Englisch- und Deutschkursen waren ihrerseits die Entwicklung des verstehenden Hörens und der Sprechfähigkeit am besten gelungen. Als Hauptursache für das Mißlingen von Sprachkursen wurde in allen Fällen die Passivität der Studenten angesehen. Ein zweiter Gegenstand der Kritik war das zu schnelle Voranschreiten des Unterrichts. Auch die eingesetzten Materialien wurden kritisiert, besonders von Englischstudenten. In allen Kursen wurde fehlende Motivierung und Aktivierung als Ursache für mißglückte Kurse angegeben.

6. Ausblick

Die vorliegende Unterschung ist die einzige ihrer Art, die den Fremdsprachenunterricht aus der Sicht der Studenten untersucht. Ihre Ergebnisse sind für folgende drei Gruppen von Bedeutung: die Lehrer der Sprachenzentren, die für Planung und Organisation des Unterrichts an Sprachenzentren Verantwortlichen sowie die für die breitere Planung des Sprachenunterrichts Zuständigen.

Die Ergebnisse der vorliegenden Untersuchung zeigen eindeutig, daß die Studenten mit ihrem Sprachstudium zufrieden sind und es als nützlich ansehen. zu diesem Ergebnis führte sowohl die Untersuchung getrennt nach Sprachen als auch die nach Ausbildungsprogrammen. Wenn man die Neuartigkeit der Sprachenzentren und die ihnen in nur beschränktem Maße zu Verfügung stehenden Ressourcen in Betracht zieht, muß man sehr zufrieden mit diesem Ergebnis sein.

Trotzdem besteht kein Grund zu der Annahme, daß der Fremdsprachenunterricht automatisch auch in der Zukunft gut läuft, wenn das bisher der Fall gewesen ist. Unter der Oberfläche einer allgemeinen Zufriedenheit waren deutliche Merkmale zu beobachten, die schon jetzt negative Reaktionen auf den Fremdsprachenunterricht zur Folge haben, die sich noch verstärken werden, wenn man sie nicht korrigiert. Der Unterricht an Sprachenzentren unterscheidet sich ja grundsätzlich vom Unterricht an Spracheninstituten darin, da das Beherrschen der Fremdsprache nicht Berufsziel ist, sondern daß die Studenten im Gegenteil u. U. recht wenig am Sprachenunterricht interessiert sind. Deshalb sind Wecken und Aufrechterhalten der Sprachstudiumsmotivation von Studenten anderer Hauptfächer eine anspruchsvolle und wichtige Aufgabe des Fremdsprachenunterrichts an Sprachenzentren und ständige Herausforderung an ihn.

Schnellste Lösung fordern einige die Unterrichtsorganisation betreffende Fragen. Als erste davon sei die Koordinierung der Sprachkurse mit dem Fachstudium genannt. Augenscheinlich ist, daß die Sprachkurse nicht zu früh oder zu spät stattfinden dürfen, da dann der Grundsatz des Sprachenunterrichts, seine Nützlichkeit für Studium und Beruf, nicht erfüllt werden kann. Man ist auf dem Holzweg, wenn Studenten des ersten Studienjahres, die ihr eigenes Fach noch nicht genügend kennen, Unterricht im fachsprachlichen Fremdsprachenunterricht erhalten oder Studenten des vierten Studienjahres, die schon mehrere Jahre fremdsprachige Bücher für ihre Prüfungen lesen mußten, im Textverstehen unterrichtet werden. Dringendste Aufgabe ist daher, ein einheitliches System zu entwickeln, das die Sprachkurse mit dem übrigen Studienprogramm optimal koordiniert.

Ebenso dringende Klärung verlangen die genaue Bestimmung der Sprachkursziele sowie die Aufklärung der von diesen Zielen abgeleiteten Bewertungskriterien. Da die Unterrichtseinheiten an Sprachenzentren kurz sind, kommt der Zielbestimmung äußerst große Bedeutung bei, um die kurze Zeit möglichst effektiv nutzen zu können. Und da gute Fremdsprachnenkenntnisse eine Voraussetzung für das Erreichen des akademischen Grades sind, muß ihrer Bewertung tatsächliche große Aufmerksamkeit geschenkt werden. Eine Vereinheitlichung der Zielbestimmungen und der Bewertungskriterien ist aus den genannten Gründen also möglichst weitgehend anzustreben.

Wie schon gesagt, sind das Wecken und Aufrechterhalten der Motivation im Fremdsprachenunterricht an Sprachenzentren wichtige Aufgaben. An die Motivation sind auch mehrere Faktoren gebunden, denen wir jetzt wachsende Aufmerksamkeit schenken müssen.

Als erstes seien die obligatorischen Sprachkurse genannt. Bisher fordert nur eine geringe Minderheit der Studenten die Umwandlung in fakultative Kurse, aber das Auftreten des Problems allein gibt schon zu denken. In der Praxis sieht die Situation an den mit fehlenden Ressourcen ringenden Sprachenzentren so aus, daß man die besten Schüler die Aufnahmeprüfung bestehen läßt, so daß nur die schwächeren an den Sprachkursen teilnehmen müssen. Solch eine Handhabung ist absurd; die Voraussetzungen zum Gelingen des Sprachunterrichts müssen garantiert werden.

Ein zweiter wichtiger Gesichtspunkt schließt sich an das Problem der obligatorischen Sprachkurse an: wird der Fremdsprachenunterricht als notwendiges Übel empfunden und man will ihn so leicht und schnell wie möglich hinter sich bringen, zerfällt die ganze Idee des Sprachenzentrenunterrichts. Die Sprachkurse sollen ja gerade auf die Bedürfnisse der Studenten zugeschnitten sein und ihnen Nutzen bringen. Wenn die Sprachenzentren nur obligatorische Kurse anbieten, bleibt der Sprachenunterricht auf halbem Wege stehen. Die Sprachenzentren haben daher die Kursauswahl zu erweitern, die den Studenten gewünschte Anfänger-, Fortgeschrittenen- und Spezialkurse bieten kann. Diese Studie zeigte deutlich, daß der fakultative Unterricht in Zukunft stärker betont werden muß. Wenn das Wecken der Motivation im Sprachenunterricht Aufgabe der Sprachenzentrenausbildung ist, muß sie auch für das Wecken der Begeisterung sorgen!

Abschließend sei festgestellt, daß die Untersuchung deutlich machte, daß sich der Einsatz für den Fremdsprachenunterricht gelohnt hat. Dem Unterricht an Sprachenzentren kommt auch in Zukunft große Bedeutung bei: geht es doch um die zu erreichenden Grundkenntnisse aller Hochschulstudenten unseres Landes. Der Stand ihrer Sprachkenntnisse und der Wille, sie weiterzuentwickeln und anzuwenden wird durch den Sprachenunterricht an Sprachenzentren geprägt.

LITERATUR

Ahlström, K-G. (1971) Korkeamman opetuksen käsikirja. Helsinki: Otava.

Aho, S. (1980) Opetuksen ja opiskelun vaikeudet kansalais- ja työväenopistoissa. Kasvatustieteiden laitos, Turun yliopisto.

Aleamoni, L.M. (1972) Illinois Course Evaluation Questionnaire (CEQ). Results Interpretation Manual. University of Illinois. Urbana, Ill.

Ausubel, D.P. (1968) Educational Psychology. A Cognitive View, New York: Holt, Rinehart and Winston.

Bendig, A.W. (1954) A factor analysis of student ratings of psychology instructors on the Purdue scale. Journal of Educational Psychology, 45, ss. 385-393.

Bendig, A.W. (1955) Ability and personality characteristics of introductory psychology instructors rated competent and empathetic by the students. Journal of Educational Research, 48, ss. 705-709.

Berk, R.E. (1979) The Construction of Rating Instruments for Faculty Evaluation. Journal of Higher Education, 50.

Brandenburg, D.C., J.A. Slinde and E.E. Battista (1977) Student Ratings of Instruction: Validity and Normative Interpretations. Research in Higher Education, 7, ss. 67-78.

Braskamp, L.A. (1980) The Role of Evaluation in Faculty Development. Studies in Higher Education, 5, ss. 45-53.

Centra, J.A. and B. Rose.(1976) Student ratings of instruction and their relationship to student learning Research Bulletin, February, Princeton, N.J.: Educational Testing Service.

Centra, J.A. (1977) Student Ratings of Instruction and Their Relationship to Student Learning. American Educational Research Journal, 14, ss. 17-24.

Clark, K.E. & R.J. Keller (1954) Student ratings of college teaching. In R.A. Eckert (ed), A University Looks at Its Program. Minneapolis: University of Minnesota Press.

Cohen, A.M., J.M. Trent and C. Rose (1973) Evaluation of Teaching. Travers, R.M.W. (ed) Second Handbook of Research on Teaching. A Project of the American Educational Research Association. Chicago, Ill.

Cohen, P.A. (1981) Student Ratings of Instruction and Student Achievement: A Meta-analysis of Multisection Validity Studies. Review of Educational Research, 51, ss. 281-309.

Costin, F., W.T. Greenough and R.J. Menges (1971) Student Ratings of College Teaching: Reliability, Validity and Usefulness. Review of Educational Research, 41. ss. 511-535.

Crooks, T.J. and M.T. Kane (1981) The Generalizability of Student Ratings of Instructors: Item Specificity and Section Effects. Research in Higher Education, 15, ss. 305-313.

Domino, G. (1971) Interactive effects of achievement orientation and teaching style on academic achievement. Journal of Educational Psychology, 62, ss. 427-431.

Dressel, P.L. (1960) "Evaluation of instruction", in Proceedings of Summer Institute on Effective Teaching for Young Engineering Teachers, at the Pennsylvania State University, Lancaster.

Ekola, J. ja T. Vaherva (1980) Aikuisopetusopas. 3. uudistettu painos. Helsinki: Tammi.

Elliot, D.H. (1950) Characteristics and relationships of various criteria of colleges and university teaching. Purdue University Studies in Higher Education 70, ss. 5-61.

Elmore, P.B. and K.A. Lapointe (1974) Effects of teacher sex and student sex on the evaluation of college instructors. Journal of Educational Psychology, 66, ss. 386-389.

Elmore, P.B. and K.A. Lapointe (1975) Effect of teacher sex, student sex, and teacher warmth on the evaluation of college instructors. Journal of Educational Psychology, 67, ss. 368-374.

Erickson, G.R. and B.L. Erickson (1979) Improving College Teaching: An Evaluation of a Teaching Consultation Procedure. Journal of Higher Education, 50.

Feldman, K.A. (1978) Course Characteristics and College Students' Ratings of Their Teachers: What We Know and What We Don't. Research in Higher Education, 9, ss. 199-242.

Feldman, K.A. (1979) Consistency and Variability among College Students in Rating Their Teachers and Courses: A Review and Analysis. Research in Higher Education, 6, ss. 223-274.

Feldman, K.A. and T.M. Newcomb (1973) The Impact of College on Students. San Fransisco: Jossly-Bass, II.

Ford, N. (1980) Levels of Understanding and the Personal Acceptans of Information in Higher Education. Studies in Higher Education, 5, ss. 63-70.

Franke-Wikberg, S. och M. Johansson (1976) Utvärdering av undervisning. Lund: Studentlitteratur.

Frey, P.W. (1973) Student Ratings of Teaching: Validity of Several Rating Factors. Science, 182.

Frey, P.W., D.W. Leonard and W. Beatty (1975) Student Ratings of Insturction: Validation Research. American Educational Research Journal, 12, ss. 435-447.

Gaff, J.G. (1973) Making a Difference: The Impacts of Faculty. Journal of Higher Education, XLIV, ss. 605-621.

Gage, N.L. (1961) The appraisal of college teaching. Journal of Higher Education, 32, ss. 17-22.

Gage, N.L. (1974) "Students' rating of college teaching: their justification and proper use", in Glasman, N.S. and Killait, B.R., eds, Second UCSB Conference on Effective Teaching. University of California at Santa Barbara, California. ss. 72-86.

Gibb, C.A. (1955) Classroom behavior of the college teacher. Educational and Psychological Measurement, 15, ss. 254-263.

Goldschmid, M.L. (1978) The Evaluation and Improvement of Teaching in Higher Education. Higher Education, 7, ss. 221-245.

Good, H.M. (1975) Instructional development, What? Why? How?. Canadian Journal of Higher Education, 5, ss. 23-51.

Good, H.M., W.H. Dowdeswell and R. Harmsen (1980) Modelling and Evaluation. Studies in Higher Education, 5, ss. 33-42.

Granzin, K.L. and J.J. Painter (1975) A multivariate analysis of factor underlying student evaluations of college instructors. California Journal of Educational Research, 26, ss. 96-106.

Greenwood, G.E., C.M. Bridges, Jr, W.B. Ware and J.E. McLean (1973) Student Evaluation of College Teaching Behaviors Instrument: A Factor Analysis. Journal of Higher Education, XLIV, ss. 598-604.

Greenwood, G.E. and H.J. Ramagli (1980) Alternatives to Student Ratings of College Teaching. Journal of Higher Education, 51.

Hogan, T.P. (1973) Similarity of student ratings across instructors, courses, and time. Research in Higher Education, 1, ss. 149-154.

Hoyt, D.P. and G.S. Howard (1978) The Evaluation of Faculty Development Programs. Research in Higher Education, 8, ss. 25-37.

Isaacson, R.L., W.J. McKeachie, J.E. Milholland, Y.G. Lin, M. Hofeller, J.W. Baerwaldt and K.L. Zinn (1964) Dimensions of student evaluations of teaching. Journal of Educational Psychology, 55, ss. 344-351.

Jones, J. (1981) Students' Model of University Teaching. Higher Education, 10, ss. 529-549.

Jussila, J. ja R. Rimõn (1974) Opiskelijoiden persoonallisuus ja opintoaste kurssiarviointien selittäjänä. Kasvatus 1974, 5, ss. 162-169.

Keaveny, T.J. and A.F. McGann (1978) Behavioural Dimensions Associated with Students' Global Ratings of College Professors. Research in Higher Education, 9, ss. 333-345.

Kohlan, R.G. (1973) A Comparison of Faculty Evaluations Early and Late in The Course. Journal of Higher Education, XLIV, ss. 587-595.

Korpimies, L. ja I. Utriainen (1981) Kielikeskusjärjestelmän resurssi- ja organisaatiokartoitus. Kielikeskusopetuksen seurantatutkimus, I vaihe. Korkeakoulujen kielikeskuksen julkaisuja N:o 14/1981. Jyväskylän yliopisto.

Kulik, J.A. and W.J. McKeachie (1975) The Evaluation of Teachers in Higher Education. Review of Educational Research, 3, ss. 210-239.

Lumsden, K.G. (1974) Efficiency in Universities: The La Paz Papers. ss. 175-203. Amsterdam: Elsevier Scientific Publishing Company.

Marsh, H.W. (1977) The Validity of Students' Evaluations: Classroom Evaluations of Instructors Independently Nominated as Best and Worst by Graduating Seniors. American Educational Research Journal, 14, ss. 441-447.

Marsh, H.W. (1980) The Influence of Student, Course, and Instructor Characteristics in Evaluations of University Teaching. American Educational Research Journal, 17, ss. 219-237.

Marsh, H.W. and J.U. Overall (1981) The Relative Influence of Course Level, Course Type, and Instructor on Students Evaluations of College Teaching. American Educational Research Journal, 18, ss. 103-112.

McKeachie, W.J. (1965) Research on Teaching at the College and University Level, teoksessa N.L. Gage (toim.) Handbook of Research on Teaching, ss. 1118-1135 ja 1157-1165.

Menges, R.J. (1973) Evaluating Learning and Teaching. San Francisco: Jossey-Baas.

Meredith, G.M. (1969) Dimensions of faculty-course evaluation. The Journal of Psychology 1969, 73, ss. 27-32.

Moen, R. and K.O. Doyle (1978) Measures of Academic Motivation: A Conceptual Review. Research in Higher Education, 8, ss. 1-23.

Naftulin, D.H., J.E. Ware Jr., and F.A. Donnelly (1973) The Doctor Fox lecture: A paradigm of educational seduction. Journal of Medical Education, 48, ss. 630-635.

Nerenz, A.G. and C.K. Knop (1982) A Time-Based Approach to the Study of Teacher Effectiveness. Modern Language Journal, 66, ss. 243-254.

O'Hanlon, J. and L. Mortensen (1980) Making Teacher Evaluation Work. Journal of Higher Education, 51.

Olkinuora, E. (1979) Oppimisen ja opiskelun mielekkyys. Katsaus kirjallisuuteen ja lähtökohtiin sekä tutkimusprojektin esittely. Kasvatustieteiden tutkimuslaitos. Selosteita ja tiedotteita n:o 121. Jyväskylä.

Palva, I.P. and V. Mikkonen (1974) Assessment of teaching by an inquiry. Med Welt 1974, 25, ss. 285-288.

Perlberg, Arye (1979) Evaluation of Instruction in Higher Education: Some Critical Issues. Higher Education, 8, ss. 141-157.

Popham, J.W. (1974) Teacher Evaluation and Domain-Referenced Measurement. Educational Technology, XIV.

Rayder, N.F. (1968) College student ratings of instructors. Journal of Experimental Education, 37, ss. 76-81.

Rees, R.D. (1969) Dimensions of students' points of view in rating college teachers. Journal of Educational Psychology, 60, ss. 476-482.

Remmers , H.H. (1968) The relationship between students' marks and students' attitudes toward instructors. School and Society, 28, ss. 759-760.

Rotem, A. (1978) The Effects of Feedback from Students to University Instructors: An Experimental Study. Research in Higher Education, 9, ss. 303-318.

Salo, M.A. ja V. Kuusela (1976) OPEVA - opintosaavutusten mittauksen ja kurssiarvioinnit yhdistävä evaluaatiosysteemi. Turun yliopiston opintoasiaintoimiston julkaisuja 2/1976. Turun yliopisto.

Seldin, P. (1975) How Colleges Evaluate Professors: Current Policies and Practices in Evaluating Classroom Teaching Performance in Liberal Arts Colleges. Croton-on-Hudson, NY: Blythe-Pennington.

Smalzreid, N.T. and H.H. Remmers (1943) A factor analysis of the Purdue Rating Scale for Instructors. Journal of Educational Psychology, 34, ss. 363-367.

Smock, R.H. and T.J. Crooks (1973) A Plan for the Comprehensive Evaluation of College Teaching. Journal of Higher Education, XLIV, ss. 577-586.

Smock, R.H. and D.C. Brandenburg (1978) Student Evaluation of Academic Programs. Journal of Higher Education, 49.

Student Instructional Report (SIR) (1971) Educational Testing Service. Princeton, New Jersey.

Stufflebeam, D.L. (1968) Toward a Science of Educational Evaluation. Educational Technology, 8, ss. 5-12.

Thorne, G.L. (1980) Student Ratings of Instructors: From Scores to Administrative Decisions. Journal of Higher Education, 51.

Vaherva, T. (1983) Koulutuksen vaikuttavuus. Käsiteanalyyttista tarkastelua ja viitekehyksen hahmottelua. Jyväskylän yliopisto. Kasvatustieteen laitoksen tutkimuksia A 1. Jyväskylä.

Witley, S.E. and K.O. Doyle (1978) Dimensions of Effective Teaching: Factors or Artifacts. Educational and Psychological Measurement, 38, ss. 107-117.

Vuori, H., S. Kokko, A. Nissinen, K. Peltonen ja T. Vaskilampi (1974) Opetuksen ja oppimisen evaluaatio. Kuopion korkeakoulun julkaisuja. Hallinto B:1. Kuopion korkeakoulu.

Heikki Nyyssönen
Oulun yliopisto

THEORETICAL ASPECTS OF THE ANALYSIS OF CONVERSATIONAL DISCOURSE

Discourse analysis (DA) means the study of text and conversation as interactive events. DA is either descriptive or applied or both. Descriptive DA is a field of linguistics, but means a considerable extension of its scope, to include structures and units which go beyond the single clause or sentence. Applied DA is a field of applied linguistics or sociolinguistics. It is mainly due to sociolinguistics, in fact, that DA has emerged in its present form.

In descriptive linguistics, discourse, in so far as it is recognised at all, is mainly treated as another level of grammar, namely the one above the sentence or, in some cases, the paragraph (Longacre 1976). Alternatively, discourse is regarded as more or less synonymous with text, defined as "a unit of language in use...any passage, spoken or written, of whatever length, that does form a unified whole" (Halliday & Hasan 1976, p. 1). In generative grammar the study of discourse is relegated to the domain of performance, ie. pragmatics. In the course of time our conceptions of performance and competence have changed and the new concept of communicative competence has emerged. At present pragmatics is rapidly developing into a systematic theory of language use (Leech 1983) and has in fact become the theoretical backbone supporting the more descriptive approach of DA.

DA in its present state is largely a British phenomenon, centred in such universities as Birmingham, Lancaster and Nottingham. It is often concerned with practical applications, in the domain of language teaching in particular (Widdowson 1979). In the United States, as well, DA is often closely associated with pedagogic concerns, especially within what is called second-language research (Larsen-Freeman 1980).

We have noted above that discourse is often equated with text. There is also a tendency to use the two notions, discourse analysis and text linguistics (TL) more or less synonymously. We can, however, think of DA and TL as two different approaches to the analysis of suprasentential stretches of language (Edmondson 1981). In Brown & Yule (1983) text is defined as the verbal record of a communicative act, and in the study of natural conversation, for instance, we must rely upon text, ie. the transcript, for the analysis of the "underlying" discourse. Apart from the spoken words, the transcript will contain information about pauses, stresses, intonation patterns, etc. Brown & Yule (1983) refer to such transcripts as spoken texts (p. 9). They exemplify natural or "live" conversational discourse. Conversational discourse may also be composed, as in plays and novels. Written texts exemplify what Pawley and Syder (1983) call "autonomous" discourse. Such discourse is created by a composer who is separate and remote from the audience. The audience is anonymous, the messages impersonal, etc. Written texts are typically monologic, pre-planned, edited and revised. Irrespective of whether the text is conversational or autonomous the focus of DA is not upon the surface features of the text but upon the interactional structure underlying the text. This structure is described in terms of functional units such as the phase, exchange, move and act (Edmondson 1981).

In Edmondson (p. 4) a discourse is defined as a structured event manifest in linguistic (and other) behaviour. A text, by contrast, is a structured sequence of linguistic expressions forming a unitary whole (cf. Halliday & Hasan's definition on previous page). The structured sequences of linguistic expressions (ie. clauses, sentences and paragraphs) that form a text are described in text linguistics (Halliday & Hasan 1976, de Beaugrande & Dressler 1981) as well as so called text grammars (eg. Werlich 1976).

Cohesion is one of the basic problems of any text-centred approach. In a discourse-centred approach it is the notion of coherence that is the most fundamental concern. Coherence may be seen as either global or local. The global coherence of a discourse is based on the fact that it is divided, in a hierarchical manner, into phases, exchanges and moves. Within the smaller units there is local coherence. As an example consider following dialogue:

(1) A: How's the thesis going?
B: I'm typing it up now - typing up the final copy.

(1) is perceived as a coherent exchange. The organisation may be described, for instance, in terms of an Opening move (A) followed by a Responding move (B), or O/R, for short. In Edmondson (1981) the two moves would be called Proffer and Satisfy. Strictly speaking (B) consists of two separate

moves: Satisfy ("I'm typing it up now") and a strategic supportive move called Expander ("typing up the final copy").

The coherence of (1) may also be described with reference to the act level. Acts are the linguistic units which *manifest* the interactional discourse structures (such as the exchange and the move). Edmondson calls them communicative acts. A communicative act is a verbal act which realises an element of interactional structure. Communicative acts are both interactional and illocutionary. As an interactional act the utterance "how's the thesis going?" realises an Opening move or Proffer in (1). Similarly the utterance "I'm typing it up now" realises a Responding move or Satisfy. Both utterances are also illocutionary acts, or illocutions, because they convey an attitude of the speaker. A's utterance conveys his wish to talk about B's thesis, while B's utterance communicates his compliance with A's wish and conveys the information which was requested by A. In Edmondson's terms A's illocution is a Request for Tell, while B's illocution is a Tell. It is in this way that the exchange produces so called Outcome (Edmondson 1981, p. 80).

According to Edmondson, illocutions are considered independently of their position in discourse structure. The illocution called Tell, for instance, is defined as follows: S(peaker) wishes H(earer) to gain information about himself and thus create or cement a social bond between himself and H. This definition allows Tell to occupy either an Opening or a Responding slot in discourse structure. In (1) it occupies the latter.

In other models, such as Burton (1981), acts are defined solely according to their internal function within the discourse itself, which depends upon their placing. Thus Reply, for instance, is defined as having the function of providing "a linguistic response appropriate to a preceding Elicitation" (Burton 1981, p. 77). Acts defined in this way may be called discourse acts (Stubbs 1983, p. 149). By contrast illocutionary acts in Edmondson's sense may concern phenomenon or state of affairs outside the discourse and are defined according to the social or psychological functions that they perform. Such acts are discourse-external. They are not, therefore, discourse acts but, for instance, *social* acts. Stubbs (p. 149) points out that social acts typically have long-term consequences: if you make me a promise today, the act may still be in force in ten years' time.

Edmondson's Tell is a social act: it gives information about the speaker and creates or cements a social bond between him and the hearer. Thus Tell has a transactional as well as an interactional function (Brown & Yule 1983).

As discourse acts only A' and B's utterances in (1) would be called Elicitation and Reply (Burton 1981). Elicitation requests a linguistic response, while a Reply provides a linguistic response which is appropriate to a preceding Elicitation.

Burton's analysis is similar to Halliday & Hasan's (1976, p. 206). However, they discuss cohesion in text, rather than coherence in discourse. Dialogue (1) exemplifies what they refer to as a "rejoinder sequence". A rejoinder is any utterance which immediately follows an utterance by a different speaker and is cohesively related to it. A rejoinder that follows a question is a response. If it answers the question it is a direct response. Dialoque (1) has the pattern Question followed by a direct Response. Halliday & Hasan's approach is text-centred, while Burton's is discourse-centred, but in spite of this difference in approach there does not seem to be a great deal of difference between their results. This seems to indicate that an analysis of discourse coherence which is based upon the notion of discourse act is not really very different from an analysis of textual cohesion.

It seems that an analysis like Burton's (1981) adds little to a study of cohesion like Halliday & Hasan's (1976). Both approaches may be contrasted with a system like that of Edmondson (1981). To repeat, in Edmondson's model the verbal acts which realise elements of interactional structure are called communicative acts. A communicative act is both interactional and illocutionary. By virtue of their *interactional* aspect communicative acts are involved in the *global* coherence of a discourse. By virtue of their *illocutionary* aspect they are involved in the *local* coherence of an exchange. Finally, as discourse-external acts they *perform some social or psychological function outside the discourse itself*. It is this last dimension, in particular, that sometimes tends to be neglected in DA, especially in those models of DA which are mainly concerned with acts as discourse acts. Such models give analyses which, although illuminative of discourse *structure*, may seem impoverished of any social and psychological *content*.

Communicative acts in Edmondson's model are multi-functional, ie. have two or more functions at the same time, only in the sense that they are both interactional and illocutionary. Normally each act is assigned just one functional label as an illocution. Other analysts (eg. Stubbs 1983) point out that a verbal act can in principle perform several functions simultaneously, on different levels of abstraction. Thus a Request, for instance, may also, on a more abstract level, perform a Challenge (ie. a criticism or accusation). In addition it is suggested by Bach & Harnish (1979) that there are also so called "collateral" acts. They are non-communicative acts that are performed simultaneously with, or instead of, some illocutionary act. An example is "kidding", ie. saying something without meaning it. In Leech (1983) similar conversational behaviour is called Banter, which is "an offensive way of being friendly"(p. 144). Thus two friends may greet one another with remarks such as "Here comes trouble!" or "Look what the cat's brought in!"

Clearly the principle of multi-functionality is essential for any attempt to arrive at an analysis which is richer in social and psychological content. This may be problematic, but no more problematic than any other description within a DA framework. There is always an important practical drawback which is expressed by Levinson (1980, p. 20) as follows: "If one looks even cursorily at a transcribed record of a conversation, it becomes immediately clear that we do not know how to assign speech acts in a non-arbitrary way". However, this problem with identifying speech acts should not lead us to abandon their investigation (Brown & Yule 1983, p. 233). What we should realise is that Speech Act theory, as it is presently formulated, does not offer the discourse analyst a way of determining *how* an utterance comes to receive a particular interpreted meaning. The interpretation of discourse is based to a large extent on a simple *principle of analogy* (Brown & Yule 1983, p. 233). We are all constrained in our interpretation by similar experience in the past. In addition to relevant past experience there is another principle of interpretation, the principle of *local interpretation* (Brown & Yule, p. 59). It instructs the hearer not to construct a context any larger than he needs to arrive at an interpretation. A simple example: if he hears someone say "Shut the door" he will look towards the nearest door available for being shut.

The principles of analogy and local interpretation should warn us against "over-problematising" DA. People do succeed in assigning functions to utterances in what seems to be a routine manner, even in ambiguous cases. Another danger in DA is that of "over-analysis", such as exaggerating certain perceived features of the interaction which may not have had significance for the participants at the time. In this respect an analysis in terms of discourse acts may well seem to present a safer option.

What follows is an alternative analysis of dialogue (1). The point of this alternative analysis is to show that the utterances of (1) may be interpreted multi-functionally and that it is likely that this multi-functionality had some significance for the participants at the time. The analysis also tries to show that in order to arrive at an interpretation we must take a close look at the linguistic expressions themselves and see whether they provide us with any interpretative clues. In the analysis that follows a close look is taken at the style of B's utterance as it is reflected in its syntax and in some of its prosodic characteristics. For convenience the dialogue is repeated here as (2) and, for B's utterance, the nucleaus and intonation are also given.

(2) A: How's the thesis going?
B: I'm typing it up NÒW - typing up the final CÓPY.
(fall) (rise)

Before we go further, consider again Edmondson's analysis. According to that (1) and (2) both consist of the following sequence of moves: Proffer/Satisfy + Expander. The Proffer is realised by an illocution called Request for Tell and the Satisfy is realised by an illocution called Tell (the definition of which has been given above). In the strategic supportive move, Expander, B "tells more" than is requested - in anticipation, perhaps, of possible further queries by A about the same topic.

We can now try to add something to Edmondson's account. For this purpose we focus first on B's utterance. As noted above, it consists of two "moves" where just one would be sufficient. The second move/clause adds little to the information in the first - clearly the fact that B is typing up the thesis now implies, or at least can imply, that he is now typing up the final copy. Thus the second clause is merely a kind of paraphrase of the first. In the syntax of spoken English, however, such paraphrases are a "speech characteristic". Although redundant in strict informational terms, they are not necessarily dysfunctional in discourse terms. In those terms a paraphrase of this kind may signal a return to

the topic, for instance. Alternatively its function may be to show emphasis, as in the present case. It seems that B wants to make the point, emphatically, that the thesis is going well and that it is in the final stages now. Thus it seems that B not only "tells more", conversationally and for the benefit of the hearer, but also at the same time supports his own interests and his scholarly reputation.

A study of the intonation patterns in B's utterance lends some support to the above analysis. We note that there is an obvious difference between the two parts, or tone units. The first ends with a fall ("I'm typing it up NÒW"). This is as expected. It is an affirmative statement, containing information that is new to the hearer. According to Brazil et al (1981), falling tone in English marks the content as new. Therefore the choice of tone is, in this case, neutral or unmarked. By contrast, the second unit ends with a rise ("typing up the final CÓPY"). This is unexpected, for the following reason. We know that the second unit is merely a paraphrase of the first. Therefore its content is already "common ground". In intonation one would expect this to be signalled by falling-rising tone as it is the function of falling-rising tone in English to mark the content of the tone unit as "part of the shared, already negotiated, common ground, occupied by the participants at a particular moment in an ongoing interaction" (Brazil et al, p. 15). For this reason B's choice of rising tone is unexpected and therefore marked. What it means or implies, in this particular case, is more difficult to determine. Whatever the meaning, it must be based on the notion of *contrast*. There is a contrast on two linguistic dimensions. First, there is a contrast, on the syntagmatic axis, between the fall in the first unit and the rise in the second. This contrast signals the difference in the informative (or transactional) function of the two tone units, namely the fact that while the content of the first is new, the content of the second is common ground. In Brazil et al (1981) both fall-rise and rise are "referring" tones, in contrast with the fall tone which is the "proclaiming" tone. However, there is a difference between the two referring tones (or r tones), in that the first, fall-rise, is non-emphatic while the latter, rise, is emphatic. Therefore there is a contrast, in B's utterance, on the paradigmatic dimension as well. This is between the rise in the second unit, which is B's actual choice, and the fall-rise that he might have chosen in its stead. It is this contrast on the paradigmatic axis which marks the second unit of B's utterance as emphatic. In other words, as the earlier analysis has already suggested, B is not, in the second unit, just repeating what he has said in the first. He is repeating it for a purpose. The study of intonation reinforces the impression that the purpose is to show emphasis.

We can try now to make the analysis more specific. We have said that the emphasis in B's utterance might indicate that he is not just supporting the interests of A, by telling him what he wanted to know, but that at the same time he is supporting his own interests, preserving his own "face" (Edmondson 1981, p. 7). Questions are, at least potentially, "face-threatening" acts (Brown & Levinson 1978) because they can raise topics that are potentially awkward or embarrassing. For this reason questions can be interpreted, by the person to whom they are addressed, as Challenges. In dialogue (2) B's behaviour suggests that A's question is experienced by him, at least to some extent, as challenging. B's reply to A's question is "verbose" and, in interactional terms, marked with a specific intonation pattern. These are clues that point to emphasis. According to Brazil et al (1981) this specific intonation pattern, the emphatic referring tone, also carries social implications. It may be interpreted as an assertion of, or claim to, social dominance. In other words, B's choice of the rise tone suggests that he has perceived A's question, to some extent, as a challenging one and that he, in his reply, somehow asserts himself against the perceived "threat". It is important to notice that this self-assertion, or self-defence, which in any case is slight, is conveyed by intonation. This means that it is conveyed indirectly, by implication. It also means that it is conveyed by means of a system which is subtle and characterised by ambiguity. For these reasons the above reading is to be taken as tentative only and the significance it imputes to the social implications of the exchange should not be exaggerated. It is possible, and even likely, that the rise tone in B's second clause carries other implications in addition to those mentioned above. Apart from the communicative contrasts (on the transactional and interactional planes) there are others provided by English intonation and their availability for interpretation adds to the ambiguity and multi-functionality of the dialogue in question.

One final comment on the above analysis. Our reading of a hint of Challenge in A's speech and a hint of Defence in B's finds some support in what is known about them and their role relation. The example is

taken from the London-Lund corpus of English conversations (Svartvik & Quirk 1980). Very little background information is in fact given. What we do know is that the two participants, A and B, are both male academics, aged 43 and 34, respectively. Thus A is the senior one. From their further conversation it appears that A, the senior academic, is being consulted by B, the junior one, on jobs, awards and publishers. It is clear from all this that the situation is, in DA terms, one of "unequal power". In this situation B's bid for "dominance", as a theoretical description of part of what goes on in (2), fits in quite well with what we have already said.

The above analysis gives at least a partial answer to Levinson's remark, (quoted above) that "we do not know how to assign speech acts in a non-arbitrary way". Any answer to this depends, of course, on what Levinson is taken to mean by "speech acts". If he is taken to mean the kind of speech acts that have been the object of classification in Speech Act theory, then it seems that Levinson's point is a valid one. It has already been pointed out that Speech Act theory does not offer us a way of determining *how* an utterance comes to receive a particular interpreted meaning. Nor does it seem that the definitions given by discourse analysts, such as Burton (1981) and Edmondson (1981), are always very helpful, either. This problem becomes all the more serious in the case of those acts which are not just discourse acts but perform other functions outside the discourse itself. Although it would be unfair to dismiss work such as Sinclair & Coulthard's (1975) which deals precisely with the level of discourse acts, exclusive concentration on that level may seem superficial and even uninteresting. In any encounter between humans there are also actions at a deeper socio-psychological level, such as challenges, defences and retreats (Labov & Fanshel 1977). In our efforts to see how such abstract acts are interpreted by the participants themselves we cannot just focus upon the surface text. Instead, we must study the text in relation to the situation, taking account of everything that it relevantly implies. In any analysis there must be a careful study of the situation. Apart from that there must be a careful study of the text itself. And what is more, there must be continual cross-reference made between the two. In the case of conversational discourse, such as dialogue (1), it is not enough, obviously, just to register the grammatical features of an utterance (such as mood) and relate those to the situation in terms of some fairly self-evident "interpretive rule". The syntax of spoken vernacular language has its own features, its "speech characteristics", which we have hardly begun to study in a systematic fashion.

The analysis of abstract functions involves not only what is said by the participants but also, and crucially, what is implied (or implicated, in the Gricean sense). In the analysis of social acts we are dealing with notions which have the status of hints, suggestions, overtones and the like. The evidence for such abstractions is to be looked for in the utterances themselves, in the situation and in all that we know about the social behaviour of fellow human beings: co-operative behaviour, face-saving, problem solving, principles of analogy and local interpretation, use of background knowledge etc. Discourse is multiply organised in interactive terms: there is global and local coherence, there are a number of different levels of meaning and function. What is more, discourse is dynamic so that the patterns and meanings are susceptible to change and revision, even cancellation, in the course of the ongoing interaction. In the case of encounters of the most trivial kind this is, no doubt, an exaggeration but then those are hardly likely to interest us anyway. We are likely to be more interested in encounters that involve some problem, where there is a goal or an outcome to be achieved and negotiated, where there is some tension and some resolution at the end. In the analysis of such encounters focussing on the most superficial features would mean missing out on the most central and most relevant aspects.

In the analysis of the deeper levels of discourse we are faced with the problem of what are to be our descriptive categories. The exchange (or interchange) has established itself as one important unit, but the internal structure of the exchange is not yet clear, or as clear as it could be (Stubbs 1983). There are different models, as we have seen. For instance, to go back to (2) for a brief moment, Burton (1981) would analyse the second part of B's utterance ("typing up the final copy") as part of an act (post-head comment), rather than as a separate move. This is because, in her model, no distinction is made between "deep" discourse-structure moves (such as Proffer and Satisfy) and "surface" strategic moves (such as Expander, Grounder and Disarmer). Nor does Burton describe such "surface" phenomena of conversational discourse as fumbling, ie. the various phrases we use to fill gaps and gain time, hedge opinions, introduce illocutionary acts etc.

One of the biggest problems in DA is still no doubt the identification of act-level units. As we have seen, there is a variety of act types. There are the discourse acts (like Burton's Elicitation and Reply), there are the illocutionary acts (like Edmondson's Tell), there are the collateral acts (like Bach & Harnish's kidding, circumlocution, small talk etc.). Finally there are the social acts, such as Challenges, Defences and Retreats. All these refer to some important aspects of conversational behaviour and, for that matter, many of them are applicable to the autonomous discourse of written texts. Obviously our decision as to which of these categories to use will depend upon our purposes and the kind of material that we work with. Illocutionary acts have perhaps received most attention. Yet there are considerable differences among the various treatments. Some analysts of conversations, especially those working within the ethnomethodological frame of reference (Sacks, Schegloff & Jefferson 1974), reject the kind of intentions that are presupposed by the notion of illocutionary act. This causes them to drop the category altogether and operate with other concepts such as the turn and the move. As regards the specific treatments of illocutions, the one in Edmondson (1981) seems promising in many respects. It has already been mentioned that he wishes to consider illocutions as independent of their placing in discourse structure. This he can do, given the kind of overall model that he has designed. As regards the labelling of acts, in particular, Edmondson uses everyday common-sense terms (such as Suggest, Propose, Complain etc.) where he finds them appropriate. Elsewhere he adopts technical labels such as Willing, Resolve, Minimisation and so on. *Willing*, for instance, is defined as follows: S(peaker) wishes H(earer) to believe that S is not against performing a future act A as in the interests of H (p. 142). This makes it possible to identify instances of Willing in those cases where we might as well apply the everyday terms "promise" or "offer". In addition, it also makes it possible to identify instances of Willing in cases such as B's reply in the following:

(3) A: Carry my bag, will you please?
B: Sure, if you like.

If we were to report what B said in everyday terms we would say, for instance, that B consented or agreed to A's request, or, perhaps, that B promised to do as requested. In Edmondson's terms what happens is that B communicates to A his willingness to perform a certain action.

Another illocution in Edmondson's system is Propose, defined as follows: S wishes H to believe that S is in favour of an act A, to be performed jointly by S and H, as in the interests of both. This makes it possible to identify both utterances in the next exchange as instances of Propose:

(4) A: Let's go to the pictures tonight.
B: Good idea, let's do that.

In such cases, "agreeing" to a "suggestion" for joint action involves making that same "suggestion". (Agree and suggest are non-technical, everyday terms.)

Commenting on his own model, Edmondson admits that it is an approach that is likely to seem counter-intuitive (p. 138). But he can also explain why it should be so: "The reason is precisely that the illocutionary terms which are available as part of the lexis of English, and which have been taken over in various approaches to speech acts, often carry strong connotations concerning the *interactional status or function* of the act so described" (my emphasis). In other words, such commonsense lexical items of English as "consent" and "agree" carry the strong connotation that any utterance labelled Consent or Agree immediately follows an utterance by a different speaker and is interactionally related to it. It is strong connotations of this kind that Edmondson wants to avoid, for his claim is that illocutions and interactional acts are to be distinctively characterised.

We cannot accuse Edmondson of inconsistency, at least in this particular case, but we can still say that the analysis remains counter-intuitive. For instance, it seems counter-intuitive to report B's utterance in (4) by saying that it was B who did the proposing. This cannot be how it is perceived by A, either. Clearly it is perceived by A that it was he who did the proposing (or suggesting) and that what B did was to agree to it. This is our intuitive analysis of the episode, but of course (as Edmondson points out) it depends upon the use of "illocutionary terms which are available as part of the lexis of English" (which is what he wants to avoid).

If our intuitions do not always agree with Edmondson's analysis on the act level, at least we know the reason. Moreover, we know that acts in his model are two-faced. If one of the faces looks unfamiliar we can turn to the other. This is the interactional aspect of the act, the fact that it realises a move in discourse structure. In (3) and (4) B's verbal

act realises a Satisfy in the exchange, ie. communicates to A that his "perlocutionary intent" has been successful - that he has succeeded in getting B to do what he wanted. This does seem intuitive. If not on the level of illocutionary structure, at least on the level of interactional structure Edmondson's analysis purports to agree with our intuitions. The difficulty seems to be that it is not easy to distinguish between these two levels. They merge into one another in our everyday thinking and in our everyday speech. Edmondson's reminder concerning the distinction between theoretical terms and pre-theoretical or non-theoretical terms may well be a salutary one. Conversation is a non-theoretical, everyday occurrence and we employ a non-theoretical, everyday metalanguage in referring to it. A model of conversational discourse is, however, a theory the purpose of which is to explain conversational behaviour by reducing such behaviour to formal concepts and rules. Such a model need not necessarily "look like" what it represents - the technical terms used by an analyst need not necessarily look like those everyday terms with which we are familiar as conversationalists.

Edmondson's model is based on a limited set of data. They consist of simulated conversations in a controlled range of situations. These are encounters between strangers so designed that they involve the negotiation of some "business" such as the achievement of mutual understanding or the restoration of social harmony. A typical situation is the following: a student returns to her books in the library to find that her seat has been taken by another. It is a stock situation, not really very different from those found in some textbooks. The language is predictable so that in the library situation what we expect and get are apologies, complaints, excuses etc. predictably distributed between the participants. It is almost as if the model had been set up first and the data generated afterwards by it. In this respect Edmondson's model is both theory-specific and data-specific at the same time. This of course restricts its applicability in other contexts.

The model also suffers from some other disadvantages. Thus some of the move types are found to be confusing in practice, especially, perhaps, Contra and Counter. The assignment of just one function per act has already been mentioned. As regards the inventory of act types there are several criticisms that can be made. Some of the labels are too general and too gross-grained (eg. Request). Others may cause confusion or overlap in practice (Tell, Claim, Opine), a few seem outlandish (Condone, License, Justify). Owing to the strictly controlled nature of the data used there is an absence of labels for acts we might want to identify in more casual encounters. Act definitions are based on such terms as "A is in favour of x", "A is not against y", "x is in the interests of B", etc. These help differentiate a particular act from the other acts within the model, but not necessarily from other acts outside the model. Again, there is specificity, but it is achieved at a cost, at the expense of comparability with other models and applicability in other contexts. Further, the encounters are between *strangers* which also affects act definitions. As we have seen, Tell, for instance, is defined in a way that restricts its scope to the kind of information only which concerns the speaker himself. Simultaneously, however, the definition confers to Tell another, *social* function (a phatic one, to be precise) in that Tell is said to create or cement a social bond between the speaker and hearer. It is in cases like this that the definition of a communicative act in Edmondson's model fuses together an illocutionary as well as a social act. If the two levels were kept separate, one could broaden the scope of Tell to cover information that need not necessarily concern the speaker himself only (cf. Report) and one could then deal with the social dimension on a different level of abstraction.

The above remarks are directed to some shortcomings in Edmondson's model for the analysis of conversational discourse. At the same time it must be acknowledged that the model represents, in many ways, an improvement on some of its "rivals". Although based upon specific data of a *formal* kind, it does contain a number of clues to the study of other, less formal encounters. In one respect, however, the model leaves something essential to be desired. This is that there is, in fact, very little information concerning the relationship between discourse and the *language* of those verbal acts by which discourse is manifested. Other models (eg. Sinclair & Coulthard 1975) at least try to explore the relationship between discourse and grammar. Stubbs (1983) contains useful observations on surface syntax and lexicon in relation to deeper interactional levels. Brazil et al (1981) investigate the discourse functions of intonation. By comparison, Edmondson is curiously uninformative in this crucial respect. In general it seems that we must await more systematic descriptions of the language vernacular discourse, covering all of its linguistic aspects,

before it will be possible to establish reliable correlations, in significant numbers, between the two levels, discourse and language. For the time being, the study of conversational discourse and the study of oral language must proceed side by side and hand in hand. Progress may seem slow, happen as it must in piecemeal fashion. But progress in linguistics is never fast and the more we find out the more we find that we must find out more.

References

Bach, K. and Harnish, R.M. 1979. Linguistic Communication and Speech Acts. Cambridge, Mass.: MIT Press.

de Beaugrande, R. and Dressler, W. 1981. Introduction to Text Linguistics. London: Longman.

Brazil, D., Coulthard, M. and Johns, C. 1981. Discourse Intonation and Language Teaching. London: Longman.

Brown, G. and Yule, G. 1983. Discourse Analysis. Cambridge: Cambridge UP.

Brown, P. and Levinson, S. 1978. Universals in language usage: Politeness phenomena. Esther N. Goody (ed.) Questions and Politeness. Cambridge: Cambridge UP.

Burton, D. 1981. Analysing spoken discourse. Malcolm Coulthard and Martin Montgomery (ed.) Studies in Discourse Analysis. London: Routledge & Kegan Paul.

Edmondson, W. 1981. Spoken Discourse. London and New York: Longman.

Halliday, M.A.K. and Hasan, R. 1976. Cohesion in English. London: Longman.

Labov, W. and Fanshel, D. 1977. Therapeutic Discourse. New York: Academic Press.

Larsen-Freeman, D. (ed.) 1980. Discourse Analysis in Second Language Research. Rowley, Mass.: Newbury House.

Leech, Geoffrey N. 1983. Principles of Pragmatics. London and New York: Longman.

Levinson, S. 1982. Pragmatics. Cambridge: Cambridge University Press.

Longacre, R. 1976. An Anatomy of Speech Notions. Lisse: The Peter de Ridder Press.

Pawley, A. and Syder, F.H. 1983. Natural selection in syntax: notes on adaptive variation and change in vernacular and literary grammar. Journal of Pragmatics. 7/5 (551-580).

Sacks, H., Schegloff, E. and Jefferson, G. 1974. A simplest systematics for the organization of turn-taking for conversation. Language 50 (696-735).

Sinclair, J. McH. and Coulthard, R.M. 1975. Towards an Analysis of Discourse. Oxford: Oxford UP.

Stubbs, M. 1983. Discourse Analysis. Oxford: Basil Blackwell.

Svartvik, J. and Quirk, R. (ed.) 1980. A Corpus of English Conversations. Lund: G.W.K. Gleerup.

Werlich, E. 1976. A Text Grammar of English. Heidelberg: Quelle & Meyer.

Widdowson, H.G. 1979. Explorations in Applied Linguistics. London: Oxford University Press.

Hartmut Schröder

Language Centre for Finnish Universities, Jyväskylä

ANMERKUNGEN ZUR PLANUNG TEXTORIENTIERTER LEHRWERKE FÜR LESEKURSE IM BEREICH DER SOZIALWISSENSCHAFTEN - ERFAHRUNGEN AUS FINNLAND - *

- Zur Bedeutung des Lesens im fachsprachlichen Fremdsprachenunterricht
- Allgemeine Grundlagen der Lehrwerkplanung
- Zielgruppenbeschreibung und Bedarfsanalyse
- Lernzielbestimmung
- Textauswahl und Textanalyse
- Aufbau und Progression
- Übungssysteme

Zusammenfassung

Der Neuorientierung der Fachsprachenforschung folgend, hat auch die fachsprachliche Lehrwerkplanung verstärkt die Textebene zu berücksichtigen: Der Fachtext hat Ausgangs- und Zielpunkt der Lehrwerkplanung und des Unterrichts selbst zu sein. Für den Bereich der Sozialwissenschaften wird ein solches Vorgehen zur Erstellung von Lehrwerken für Lesekurse diskutiert und Erfahrungen eines Projekts in Finnland berichtet. Insbesondere werden Konsequenzen für die Zielgruppenbeschreibung, Bedarfsanalyse, Lernzielbestimmung, Textauswahl, für den Aufbau von Lehrwerken und für die Erstellung von Übungssystemen aufgezeigt. Interdisziplinäre Zusammenarbeit und sorgfältige Textanalysen werden als Voraussetzungen eines textorientierten Vorgehens genannt.

* Der Artikel erscheint in erweiterter Fassung auch in der Zeitschrift "Fachsprache".

Hartmut Schröder
Language Centre for Finnish Universities, Jyväskylä

ANMERKUNGEN ZUR PLANUNG TEXTORIENTIERTER LEHRWERKE FÜR LESEKURSE IM BEREICH DER SOZIALWISSENSCHAFTEN - ERFAHRUNGEN AUS FINNLAND -

Zur Bedeutung des Lesens im fachsprachlichen Fremdsprachenunterricht

Fachsprachlicher Fremdsprachenunterricht (FFSU) ist als vor allem "berufsqualifizierender Akt" (BEIER/MÖHN 1981) zu verstehen, der den "Zugang zu fremdsprachlichen Quellen aller Art" (NAUCK 1983) vermittelt und somit "das Partizipieren am Informationsfluß erst ermöglicht". Daß das Lesen fremdsprachiger Fachliteratur in diesem Zusammenhang eine besondere Rolle spielt und mittlerweile das "wichtigste Lernziel für die meisten Lerner überhaupt" (KARAJOLI 1981) [1] sein dürfte, ist nicht weiter verwunderlich, wenn man bedenkt, daß heute in fast allen Wissenschaftszweigen das immense Wachstum der Fachliteratur und die ständige Zunahme internationaler Kontakte relativ parallel verlaufen [2]. Daß schließlich auch Deutsch im Rahmen dieser Entwicklung "Vehikel des Technologie- und Know-how-Transfers" (BECKER 1981) geworden ist, mag erklären, warum das Lesen von Fachliteratur "zu den Hauptzielen des Deutschunterrichts im Ausland zählt" (BERNSTEIN 1981)[3]. Gegenläufig zu den oft zitierten Internationalisierungstendenzen im Wissenschaftsbetrieb stellen wir zur Zeit eine "Entwicklung eines immer breiter werdenden Spektrums von Weltsprachen" (KARAJOLI 1981) fest, von der insbesondere auch das Deutsche zu profitieren scheint.

---

1 NAUCK (1983: 116) führt Forschungsergebnisse an, denen zufolge "sich menschliches Sprachverhalten im Verhältnis 8:7:4:2 in Bezug auf die Sprachverwendungsgebiete Hörverstehen, Leseverstehen, Sprechen und Schreiben gliedert".

2 FLUCK (1980: 134) zitiert UNESCO-Statistiken, wonach 50 % der naturwissenschaftlichen Fachliteratur in Sprachen veröffentlicht werden, die mehr als die Hälfte aller Wissenschaftler der Erde nicht lesen können.

3 Vgl. auch SPITZBARDT (1983), der darauf hinweist, daß 50 % der indischen Studenten Deutsch studieren, um Zugang zu deutschsprachigen Informationsquellen zu erhalten.

Der Bedeutungszuwachs der Lesefertigkeit im Fremdsprachenunterricht manifestiert sich u.a. auch in der Gründung einer eigens für diesen speziellen Zweck vorgesehenen Zeitschrift: READING IN A FOREIGN LANGUAGE (ab 1983) reflektiert "a rapidly growing, world-wide interest in learning to read and in teaching reading in a foreign language", wie die Herausgeber im Editorial der ersten Ausgabe betonen. Was den engeren Bereich des FFSU betrifft, so wurde die Problematik bereits 1982 durch das internationale LSP-Symposium "Reading for Professional Purposes in Native and Foreign Languages" aufgegriffen [4] und ist seither ein wichtiger Forschungsbereich sowohl der Fachsprachenlinguistik, als auch -didaktik geworden.

Für Deutsch als Fremdsprache (DaF) sind in den letzten Jahren mehrere Lehrwerke entstanden, die ausschließlich auf die Entwicklung des fachsprachlichen Leseverstehens gerichtet sind [5]. Dabei fällt auf, daß die meisten Materialien dieser Art für die Naturwissenschaften und Technischen Wissenschaften vorgesehen sind, die Geistes- und Sozialwissenschaften aber weithin unbearbeitet blieben [6]. Gleichwohl scheint es einen ausgesprochenen Bedarf an Lesekursmaterialien für den Bereich der Sozialwissenschaften zu geben, was wir durch eigene Erfahrungen und Berichte aus der Praxis in anderen Ländern [7] belegen können.

Im folgenden wollen wir uns gerade auf diesen wenig bearbeiteten Bereich [8] beziehen und Möglichkeiten für die Lehrwerkerstellung aufzeigen. Verzichten müssen wir dabei auf eine Auseinandersetzung mit Vorbehalten gegenüber "reinen" Lesekursen (z.B. SCHLEUSENER 1982) und Zweifeln, ob eine Lehrwerkproduktion für diesen bestimmten Zweck überhaupt sinnvoll ist (z.B. AHRENS 1981) [9]. Gewisse pragmatische Erwägungen, auf die hier nicht weiter eingegangen werden soll, veranlassen uns, die Produktion von Lehrmaterialien für "reine" Lesekurse als eine effektive Maßnahme für den FFSU im Ausland zu verstehen.

---

4 Vgl. dazu den Konferenzbericht in: Reading in a Foreign Language, Vol. 1, No. 1, March 1983, S. 65 ff.

5 Vgl. dazu Arbeitsmittel für den Deutschunterricht an Ausländer. Herausgegeben vom Goethe-Institut, München 1984.

6 Auf die "Einseitigkeit zugunsten von Naturwissenschaften und Technik" weist seitens der Fachsprachenforschung auch HOFFMANN 1976: 98) hin.

7 Vgl. z.B. GROSSO (1981), HERRMANN (1981), ARMALEO-POPPER (1976), NEUF-MÜNKEL (1982) und NARDON (1983).

8 Vorarbeiten sind recht spärlich. Vgl. aber UESSELER (1982), ROTHACKER (1960), KUMMER (1983) und BARTEN (1983/84).

9 AHRENS (1981) schlägt vor, für jeden Lesekurs neue Materialien zu erstellen, so daß die Zielgruppen und ihr Bedarf maximal berücksichtigt werden.

Allgemeine Grundlagen der Lehrwerkplanung

In ihren "Vorüberlegungen zu einem Hamburger Gutachten" beschreiben BEIER/ MÖHN (1981) das allgemeine "Bedingungsgefüge" des FFSU und vermitteln potentiellen Lehrwerkproduzenten wichtige Kriterien zur Lehrwerkanalyse und -beschreibung. Da "Fachsprache in ihrer Totalität" nicht Gegenstand der Lehrwerkplanung (und des Unterrichts selbst) sein kann, werden die Bedürfnisse der Lernenden und der gesellschaftliche Bedarf als denkbare Entscheidungsgrundlagen angenommen [10]. Als weitere Entscheidungsebenen für die Erstellung fachsprachlicher Lehrwerke werden genannt:

1. Die Entwicklung und Begründung von Lernzielen,
2. die Beziehung von fachlichen Inhalten auf die Lernziele, die Stoffauswahl und die Bestimmung der Grobabfolge,
3. die Planung der Unterrichtsverfahren,
4. die Entwicklung entsprechender Übungs- und Kontrollformen.

BECKER (1974) schlägt zur Konzipierung von Lesekursen die Berücksichtigung folgender Ebenen vor: 1. Die Kontrastsprache (bezogen auf die Adressaten), 2. die Textsorten, 3. die Unterrichtsgestaltung (Übungsunterricht statt Übersetzungsunterricht!), 4. Fertigkeitstreue (Leseverstehen, nicht Übersetzen!), 5. fachliche Differenzierung und 6. die Textgestaltung (Originaltexte!)

Zur Orientierung von Lehrbuchautoren scheinen uns diese Hinweise (vor allem auch die recht umfangreichen Kriterienraster zur Lehrwerkanalyse und -beurteilung von BECKER 1981 und BUHLMANN 1982 a) recht nützlich zu sein, da sie dazu zwingen, die eigene Konzeption kritisch zu überprüfen und evtl. zusätzliche Entscheidungsebenen in die Planung und Produktion einzubeziehen. Wir verzichten hier jedoch auf eine weitere Beschäftigung mit diesen und anderen Ansätzen [11] und verweisen stattdessen auf JANDA/SPRISSLER (1977) und SPRISSLER (1977), die u.E. in ihrem Modell die wichtigsten Phasen zusammengefaßt haben, die jede Lehrwerkerstellung durchlaufen sollte:

1. Zielgruppenbeschreibung und Bedarfsanalyse (Kommunikationssituationen, Sprachfähigkeiten, berufliche Tätigkeit usw.),
2. Erstellung von Textreihen (Corpus) ausgehend von der Sprachverwendungssituation,

[10] Vgl. auch BUHLMANN (1981 b), die FFSU als "extrem adressatenabhängig" auffaßt; ähnlich auch EHNERT (1976).

[11] Verwiesen sei aber noch auf KÖHLER (1980), der in seinem grundlegenden Beitrag neun Bedingungen für die Konzipierung fachsprachlicher Lehrwerke nennt.

3. Linguistische Analyse des Corpus (lexikalische Inventare, Sprachverwendungsgrammatiken usw.),
4. Erstellung von Lehr- und Lernmaterialien,
5. Unterrichtsdurchführung und Evaluation.

Bevor wir uns nun im folgenden auf der Grundlage dieser fünf Phasen Fragen der Lehrwerkplanung zuwenden und Erfahrungen aus unserem Projekt "Deutsch für Sozialwissenschaftler" [12] berichten, sei vorab ein kurzer Blick in die fachsprachendidaktische und -linguistische Diskussion geworfen, da diese die beiden wichtigsten Bezugspunkte der Lehrwerkplanung sind und sich deren Entwicklung in gewisser Weise auch in der Art der fachsprachlichen Lehrwerke spiegelt.

Nachdem die Fachsprachenforschung sich zunächst mit Fragen der Lexik von Fachsprachen, sodann mit ihren syntaktischen Merkmalen beschäftigte, scheint nun eine neue Phase begonnen zu haben, die MÖHN (1983) als eine "intensive Zuwendung zum Problem Fachtext" beschreibt. Diese "Neuorientierung" (SCHLIEBEN-LANGE 1983) [13], die mit der allgemeinen Entwicklung der Linguistik verbunden zu sein scheint, macht sich auch im FFSU bemerkbar [14], und es sind für die Lehrwerkplanung bereits nützliche Anregungen gegeben worden, auf die wir uns im folgenden beziehen werden. Die "Zuwendung zum Problem Fachtext" [15] ist u.E. eine ausgesprochene Chance für die fachsprachliche Lehrwerkplanung, und wir meinen, daß sie auf allen Entscheidungsebenen zu erfolgen hätte, also Grundlage bei der Lernzielbestimmung, Textauswahl

[12] Bei dem genannten Projekt handelt es sich um ein Gemeinschaftsprojekt des Goethe-Instituts und des Zentralen Spracheninstituts der finnischen Hochschulen; vgl. dazu auch SCHRÖDER (1984).

[13] "Neuorientierung" scheint uns indes nicht ganz korrekt zu sein, da insbesondere in der Praxis des FFSU die Textebene schon seit langem eine Rolle spielt. So z.B. in der DDR bei SCHILLING (1973 !) und KAMPRAD (1975), der die Bedeutung der Absatzstruktur in gesellschaftswissenschaftlichen Fachtexten für den fremdsprachlichen Leseunterricht untersuchte, GLÄSER (1978), die den Fachtext "als den eigentlichen Schnittpunkt" sieht und in einen "größeren funktionalen Bezugsrahmen" zu bringen versucht. In der BRD beschäftigt sich BUHLMANN mit der Textebene und entwickelt spezielle Übungssysteme. NEUF-MÜNKEL (1982) berichtet, daß es zu ihrer Unterrichtspraxis gehört, die Studenten dazu zu orientieren "von oben nach unten" zu lesen und die "Makrostruktur" zu beachten.

[14] Siehe dazu BEIER/MÖHN (1984) und HOFFMANN (1983 und 1984 b).

[15] Wir verstehen unter Hinwendung zum Text nicht nur die Einbeziehung der Makrostruktur von Texten, sondern fassen darunter auch die Berücksichtigung der "Denk- und Mitteilungsstrukturen" (BUHLMANN 1982) der Fachdisziplinen. Es bleibt freilich zu erwähnen, daß in der gegenwärtigen Fachsprachenforschung eine solche "kommunikativ-funktionale Betrachtung von Fachsprachen gerade erst begonnen" hat (MATTUSCH 1984).

und der Bestimmung der Übungstypologie und Unterrichtsverfahren sein sollte: Nur durch die Berücksichtigung der Ebenen, die über Lexik und Syntax hinausgehen und pragmatische Aspekte mit einbeziehen, ist der eingangs erwähnte Anspruch des FFSU, "berufsqualifizierender Akt" zu sein und "Zugang zu fremdsprachigen Quellen aller Art" verschaffen zu können, einzulösen [16].

Zielgruppenbeschreibung und Bedarfsanalyse

Am Anfang jeder Lehrwerkplanung stehen selbstverständlich Reflexionen über die Zielgruppe/n und deren speziellen Fremdsprachenbedarf [17]. Dieser ist so genau wie möglich zu analysieren (BEIER/MÖHN 1984), da er Grundlage für die weiteren Entscheidungen ist. Erfolgen kann eine Bedarfsanalyse durch Befragung von Adressaten der Sprachlehrveranstaltungen und/oder von sachkompetenten Institutionen (Ausbildungsinstanzen, Beschäftigungsbereiche usw.). Da Befragungen von Adressaten meistens recht aufwendig sind [18], erscheint es im allgemeinen sinnvoller zu sein, Fachleute zu Rat zu ziehen, die über den speziellen Bedarf der Zielgruppe informiert sind. Mitunter gibt es aber auch schriftliche Bestimmungen (Studienordnungen u.ä.) über Fremdsprachenkenntnisse, die von den Adressaten verlangt werden.

Für die Lehrwerkplanung sollte vor allem die Erhebung und anschließende Berücksichtigung folgender Daten von Bedeutung sein:

1. Angaben über die Zielgruppe (Alter, Fremdsprachenkenntnisse, Sachkenntnisse im Fach, "study skills" und "reference skills"[19] usw.) [20],
2. Angaben über den zu erteilenden Unterricht, soweit dieser reglementiert ist und nicht der Beeinflussung der Lehrwerkproduzenten unterliegt (Zeitpunkt des Unterrichts: am Ende oder am Anfang des Studiums?, Dauer, Art, Ort, Form: freiwillig/obligatorisch?, Sprach- und Fachkompetenz der Lehrer),

16 Vgl. für den anglo-amerikanischen Sprachraum z.B. die neueren Beiträge in Reading in a Foreign Language, insbesondere WILLIAMS (1983), STANLEY (1984) und JOHN/DAVIES (1983); für die Diskussion in der Sowjetunion vor allem METS/MITROFANOVA/ODINZOVA (1981). Gleiche Tendenzen belegt SELLE (1983) auch für die französische Fachsprachenforschung.

17 Wir verzichten hier auf die Unterscheidung nach objektivem und subjektivem Bedarf, wie es jüngst der Internationale Deutschlehrerverband in einer Befragung macht.

18 Ein "Fragebogen zur Ermittlung des fachspezifischen Sprachlernbedarfs" findet sich bei NÜBOLD (1980); KEISER (1982) referiert Ergebnisse einer ausführlichen Befragung in der Schweiz. Beide Befragungen bestätigen nochmals die Bedeutung des Leseverstehens.

19 Wir verwenden diese Begriffe im Sinne von BUHLMANN (1982b)

20 Hierzu gehört grundsätzlich auch die Angabe über die Muttersprache, die in unserem Fall aber nicht erhoben werden muß.

3. Angaben über den Anwendungsbereich und mögliche Kommunikationssituationen, auf die der Unterricht vorbereiten soll (Lesen der Fachliteratur während des Studiums, Praktika im Ausland, Teilnahme an Konferenzen, Kontakte mit deutschen Fachkollegen usw.),
4. Angaben über Themen und Textsorten (Gegenstände, Sachverhalte, Strukturen und Situationen des Faches; Textsorten, die für die Adressaten von besonderer Bedeutung sind),
5. Angaben über die Art der zu vermittelnden Fertigkeiten (Hören, Lesen, Sprechen, Schreiben).

Durch die Erhebung dieser Daten bei unserer Zielgruppe erhielten wir folgendes Bild von unseren Adressaten:

- Es handelt sich um finnischsprachige [21] Studenten der Sozialwissenschaften, die aus der Schule zwei bis fünf Jahre Deutsch mitbringen, am Anfang des Studiums stehen, nur über geringe Sachkenntnisse im Fach verfügen und "study skills" und "reference skills" nur unzureichend entwickelt haben;
- der Unterricht ist stark reglementiert, findet zumeist zu Beginn des Studiums statt, ist obligatorisch oder unterliegt der Wahlpflicht; in der Regel stehen 56 Stunden im Studienhalbjahr zur Verfügung; die Lehrer sind finnische Muttersprachler und haben keine oder nur sehr geringe Vorkenntnisse in den Sozialwissenschaften [22];
- die Anwendungsbereiche sind nur schwer zu ermitteln und differieren im Einzelfall stark; gemeinsam ist allen Adressaten nur der Anwendungsbereich "Lesen der Fachliteratur während des Studiums";
- die Themen ergeben sich aus den jeweiligen Fächern der Studenten. Die Textsorten reichen vom Lehrbuch bis zum Zeitschriftenaufsatz, Wörterbuchartikel und der wissenschaftlichen Monographie;
- an Fertigkeiten soll zunächst nur das Lesen vermittelt werden; die Entwicklung weiterer Fertigkeiten erfolgt in späteren Kursen [23].

21 Von den wenigen schwedischsprachigen Muttersprachlern können wir absehen, da diese meistens zumindest als Zweitsprache Finnisch haben.

22 Zu diesem Punkt lassen sich nur schwer allgemeinere Aussagen machen, da die Situation fast an jeder Hochschule anders ist.

23 Wir sind uns der Problematik einer solchen Trennung bewußt, müssen andererseits aber als Lehrwerkproduzenten zunächst von dieser Trennung ausgehen, da das genannte Bedingungsgefüge nicht unmittelbar unserer Entscheidung unterliegt.

Lernzielbestimmung

Nachdem die Zielgruppe/n und deren spezieller Bedarf an Sprachkenntnissen beschrieben sind, lassen sich daraus Lernziele für Sprachlehrveranstaltungen und Lehrwerke ableiten. Auszugehen ist dabei von den allgemeinen Zielen, die die Fachsprachendidaktik [24] begründet, wobei gemäß dem Stand der aktuellen Diskussion die Textebene besonderer Berücksichtigung bedarf.

Nach BUHLMANN (1982b) ist das Hauptlernziel des FFSU die Vermittlung "sprachlicher Handlungsfähigkeit im Fach", d.h. der FFSU soll:
- "die im Fach gängigen Denk- und Mitteilungsstrukturen bewußt oder nachvollziehbar machen"
- "die nötigen lexikalischen und syntaktischen Mittel vermitteln"
- "die im Fach gängigen Textbaupläne (Textabläufe) verfügbar machen" [25].

Bezogen auf den fremdsprachigen Leseunterricht nennt BUHLMANN als oberstes Lernziel "Strategien zu einer funktional-kommunikativen Auseinandersetzung mit Texten zu entwickeln" und auf die "semantische Steuerung" des Verstehensprozesses zu orientieren [26]. Die Textebene findet dabei

24 Die Begriffe Fachsprachendidaktik und Didaktik der Fachsprache/n werden in der Literatur nicht einheitlich gebraucht. VON HAHN (1981) vermißt "eine fundierte allgemeine Fachsprachendidaktik" und FLUCK 1983/84) behauptet, "daß es eine Didaktik der Fachsprache bisher nicht gibt". Für begriffliche Klarheit sorgt KÖHLER (1985), für den die Fachsprachendidaktik eine "spezielle Modifikation der Didaktik des Fremdsprachenunterrichts" ist. Die Didaktik befaßt sich dabei - so KÖHLER - "mit den allgemeinen Gesetzmäßigkeiten des institutionellen Lehrens und Lernens von Fremdsprachen" und erarbeitet die "Grundlagen für die Zielkonzeption, für die Stoffauswahl und für Gestaltung der Unterrichtsprozesse". In Abgrenzung dazu untersucht und beschreibt die Methodik einer Einzelsprache "die optimalen Formen der Vermittlung und Aneignung einer bestimmten Sprache in einem streng determinierten Fremdsprachenunterricht und dient der unmittelbaren Unterrichtsgestaltung". Eine Didaktik der Fachsprachen liegt nach KÖHLER zwischen der Didaktik des Fremdsprachenunterrichts und der Methodik der Sprache Lx.

25 Eine konsequente Einbeziehung der Textebene befürworten auch SPILLNER (1983): "Aufgabe des Fachsprachenunterrichts ist weniger die Aneignung von Terminologie als viel eher die Vermittlung fachspezifischer Wortbildungsmuster, Satzschemata, Stilkonventionen, Argumentationsstrukturen etc." und KUMMER (1983): "es gilt vielmehr, sprachliche Zugänge zu den Stoffen und den ausgedrückten Ideen zu schaffen... Daneben sind die Formen wissenschaftlichen Arbeitens einzuführen".

26 Auf Probleme der Entwicklung des "verstehenden Lesens" können wir in diesem Zusammenhang nicht weiter eingehen; wir verweisen stattdessen auf die umfangreiche Literatur zu diesem Bereich: BUHLMANN (1981 a und b), KARAJOLI (1981), LÖSCHMANN/PETZSCHLER (1976), SELTMANN (1978) und ZIMMERMANN (1984). Für den FFSU sind auch die russischsprachigen Beiträge in dem Sammelband der AKADEMIJA NAUK (1975) von Interesse. Neuerdings sei besonders auf die Beiträge in Reading in a Foreign Language hingewiesen.

besondere Berücksichtigung: "Die Lerner müssen mit Techniken zur inhaltlichen Entschlüsselung von Texten ausgerüstet werden, die sich nicht an formalgrammatischen Kriterien, sondern an Kriterien des Textinhalts und des Textaufbaus orientieren" (1978: 170).

Unter Einbeziehung dieser allgemeinen Ziele der Fachsprachendidaktik setzen wir als Hauptlernziel unseres Lehrwerks "Deutsch für Sozialwissenschaftler": Die Arbeit mit sozialwissenschaftlichen Texten auf der Grundlage der Interessen der Studenten und unter Berücksichtigung der Studien- und späteren Berufsanforderungen. Teillernziele ergeben sich aus den Besonderheiten der Zielgruppe und deren Bedarf:
- Für das Studium und das fremdsprachige Lesen sollen notwendige "study skills" und "reference skills" aufgebaut werden
- die Studenten sollen den wissenschaftlichen Diskurs sozialwissenschaftlicher Texte kennen und erkennen können
- die Studenten sollen grammatische Strukturen in ihrer Funktion für den Textinhalt kennen und erkennen können
- die Studenten sollen über ein "lexikalisches Minimum" (HOFFMANN 1984a) von Struktur- und Funktionswörtern, sowie von sozialwissenschaftlichen und fachübergreifenden Schlüsselwörtern verfügen
- die Studenten sollen unbekannte Lexik mit Hilfe von Wortbildungsregeln, Vorwissen und unter Berücksichtigung des Kontextes entschlüsseln und ein- und zweisprachige Wörterbücher ökonomisch nutzen können.

Die Vermittlung der bei BUHLMANN genannten Denk- und Mitteilungsstrukturen wirkt in all diese Lernziele mit hinein und ist von zentraler Bedeutung für das gesamte Lehrwerk; denn mit AHRENS (1981) gehen wir davon aus, daß: "Wer die Prozesse wissenschaftlicher Kommunikation nicht kennt, kann diese nicht verstehen, selbst wenn die meisten Vokabeln gegeben sind" (159). Wir versuchen daher, dem Studenten "Methoden zur Identifizierung der Mitteilungsstrukturen" zu vermitteln, die ihn "im Prozeß der Entwicklung seiner Lesefähigkeit zu einer aktiven geistigen sprachlichen Handlung" veranlassen (ORTMANN/STERNAGEL 1978: 495/96) [27].

27 Der Begriff Mitteilungsstrukturen wird von ORTMANN/STERNAGEL verstanden als "Sinnzusammenhang zwischen voraufgehenden oder nachfolgenden Textabschnitten", die "die übergreifenden Sinnzusammenhänge von Textelementen bzw. Satzfolgen in Fachtexten" charakterisieren. Die Mitteilungsstrukturen unterscheiden sich in den einzelnen Wissenschaften und korrelieren mit bestimmten "fachtypischen Mitteilungsgehalten". Als Lernziel für den FFSU folgern ORTMANN/STERNAGEL, daß dem Lernenden bewußt gemacht werden soll, "daß Texte seiner Fachrichtung bestimmte fachtypische Mitteilungsgehalte realisieren, die immer wieder anzutreffen sind und entspre-

Abschließend bleibt zur Lernzielbestimmung zu ergänzen, daß die Intensität der Realisierung der recht umfangreichen und anspruchsvollen Lernziele immer im Kontext des gesamten Bedingungsgefüges des Unterrichts zu relativieren ist: Letztlich kann bei einem zeitlich begrenzten Unterricht von nur 56 Stunden eigentlich nur "Hilfe zur Selbsthilfe" (POETZELBERGER 1983) gegeben werden, durch die dem Studenten die Angst vor dem Lesen fremdsprachiger Texte genommen und eine ökonomische Lesehaltung entwickelt wird.

Textauswahl und Textanalyse

Die Auswahl der Texte erfolgt auf der Grundlage der Zielgruppenbeschreibung und Bedarfsanalyse und steht in einem funktionalen Verhältnis zu den festgelegten Lernzielen: "Die Texte und Textsorten müssen nach den Zielen des Kurses ausgewählt werden und nicht nach dem Inhalt einer deutschen Grammatik" (AHRENS 1981: 154).

Eine Textauswahl, die die eingangs genannte "Hinwendung zum Fachtext" berücksichtigt, hat insbesondere davon auszugehen, daß im FFSU Inhalte die Sprache verständlich machen und nicht umgekehrt (BUHLMANN 1983). "Primäres Auswahlkriterium" sind daher "pragmatische Kategorien", wie z.B. inhaltliche Kohärenz, Gliederung, Textsorten usw. (WILLE 1978). Will die Textauswahl schließlich "wissenschaftlich begründeten exakten Kriterien folgen" (KOWALKE 1973: 117), so sind genauere Recherchen [28] zu erstellen, da der Lehrer auf sich gestellt hier in einer "Überforderungssituation" steht, aus der ihm nur die interdisziplinäre Zusammenarbeit mit Fachleuten des zu vermittelnden Fachgebietes hilft [29]. Doch leisten die Fachleute zunächst nur eine Vorauswahl, da die Endauswahl nach "sprachlichen Gesichtspunkten" [30] zu erfolgen hat (EBERMANN 1973).

chend dem Grad der Standardisiertheit der Darstellungsweise an bestimmten sprachlichen Mitteln zu erkennen sind, an der Lexik, z.B. Nomen, Verben, Form des Prädikats usw." (s. 495).

28 KOWALKE fordert Recherchen durch Befragung von Sachkennern in drei Phasen: 1. Adressaten- und Fachgebietsrecherche, 2. Themenrecherche und 3. Textrecherchen.

29 Nach BUHLMANN (1983) hat diese zur Auswahl der Lerninhalte, zur Festlegung der Progression und zur Korrektur der Lehrwerke zu erfolgen. Die Notwendigkeit der Zusammenarbeit betonen auch BEIER/MÖHN (1984) und SPRISSLER (1977).

30 Das bedeutet nicht ein Zurück zur Lexik- und Grammatikorientierung, sondern gerade die besondere Berücksichtigung der Textebene: Enthalten die Texte die relevanten Mitteilungsstrukturen und fachspezifischen Mitteilungsgehalte usw. usf.? Das sind Fragen, die von Fachwissenschaftlern alleine oftmals nicht beantwortet werden können, da sie für diese zu wenig sensibilisiert sind.

Schematisch dargestellt läßt sich die Textauswahl auf fünf Ebenen beschreiben, denen jeweils wieder bestimmte Kriterien zugeordnet werden können [31]:

EBENE	KRITERIEN	AUSWAHL
1. Fachlich-inhaltliche Ebene	- typische Sachverhalte, Gegenstände und Situationen des Faches - angemessener Spezialisierungsgrad (Wissensebene der Studenten berücksichtigend und durch neue Informationen motivierend) - nützlich im Anwendungsbereich	durch Fachwissenschaftler der Ausgangssprache
2. Textlinguistische Ebene	- typische und im Anwendungsbereich nützliche Textsorten - geeignet, um Mitteilungsstrukturen bewußt zu machen - verständniserleichternde Kommunikationsverfahren	durch Zusammenarbeit von Lehrwerkproduzenten und zielsprachigen Fachwissenschaftlern
3. Lexikalische Ebene	- ausreichender Anteil hochfrequenter allgemeinsprachlicher und fachübergreifender Lexik - typische Fachlexik im Verhältnis zu Themen, Situationen usw.	Lehrwerkproduzenten durch Zusammenarbeit wie 2.
4. Grammatische Ebene	- ausreichender Anteil hochfrequenter Strukturen im Verhältnis zu Kommunikationsverfahren	Lehrwerkproduzenten

31 In das Schema fließen Überlegungen der zitierten Autoren ein. Vgl. außerdem BABAJLOVA/KICAEVA (1984), KAMPRAD (1971) und KUMMER (1983).

5. Ebene der Verständlichkeit	- die vier Dimensionen der Verständlichkeit (LANGER/SCHULZ V. THUN/TUSCH 1974) - geeignet, um Lesestile zu entwickeln	Lehrwerkproduzenten

Nachdem also die Vorauswahl auf der ersten Ebene durch die Fachleute getroffen wird [32], setzt die Endauswahl der Texte auf der zweiten bis fünften Ebene eine genaue Textanalyse voraus. Dieser sollte genügend Zeit und Aufmerksamkeit gewidmet werden, da in einem FFSU, der von Fachtexten ausgeht (und sie nicht nur als Medium zur Vermittlung von Lexik und/oder Grammatik einsetzt) von der Auswahl dieser die Qualität des Lehrwerks überhaupt abhängt [33].

Zur Durchführung von Textanalysen für Zwecke im Rahmen der Lehrwerkproduktion erscheinen uns insbesondere der Ansatz der "kumulativen Textanalyse" von HOFFMANN (1983) und die kommunikativ-funktionale [34] bzw. funktional-kommunikative [35] Sprachbeschreibung anwendbar zu sein, auf die wir im folgenden jedoch nicht weiter eingehen werden [36].

Unerwähnt bleiben soll zum Schluß nicht, daß die interdisziplinäre Zusammenarbeit, d.h. die Einbeziehung von Fachwissenschaftlern in den Prozeß der Textauswahl und -analyse, nicht immer ganz unkompliziert verläuft. Unsere Erfahrungen zeigen, daß es bisweilen schwerfällt, eine gemeinsame Sprache zu finden: Die Fachwissenschaftler sind nur unzureichend für sprachliche Aspekte, die Lehrwerkproduzenten nur unzureichend für fachlich-inhaltliche Aspekte sensibilisiert. Dennoch meinen wir, daß das vorgeschlagene Verfahren der Textauswahl nützlich und möglich ist und im Rahmen des beschriebenen Bedingungsgefüges die erwünschten Ergebnisse bringt [37].

32 Wir ließen von mehreren Fachleuten (unabhängig voneinander) insgesamt etwa zwei- bis dreimal soviel Texte vorschlagen, wie für das Lehrwerk benötigt wurden.

33 Nach unseren Erfahrungen beanspruchen Textauswahl und Textanalyse mitunter mehr Zeit als die eigentliche Erstellung der Lehrwerke.

34 Zur Einführung in die kommunikativ-funktionale Sprachbetrachtung vgl. BOECK/KIRSTEN (1975), als Beispiel einer solchen Textanalyse MATTUSCH (1984).

35 Eine Übersicht über die funktional-kommunikative Sprachbeschreibung findet sich bei SCHMIDT (1980), ein Beispiel einer solchen Analyse von Fachtexten bei GRAWERT (1982).

36 Beispiele für Textanalysen zu didaktischen Zwecken finden sich außerdem bei BEISBART/DOBNIG-JULCH/EROMS/KOSS (1976); vgl. grundsätzlich zur Analyse von Fachtexten auch ANDRĂ (1982) und PETÖFI 1981).

Aufbau und Progression

Durch die Textauswahl und -analyse allein entsteht noch kein fertiges, in ein Lehrwerk einzubringendes, Curriculum. Die ausgewählten Texte müssen in eine bestimmte Reihenfolge gebracht werden, die sich aus der Beschreibung der Zielgruppe und deren Sprachbedarf sowie aus grundsätzlichen didaktischen Überlegungen ergeben sollte. Was letztere betrifft, so verbietet sich eine - wie auch immer geartete - linguistische Progression im FFSU. Bleiben als Möglichkeiten das "thematische Gliederungsprinzip" und der Aufbau nach Kommunikationsverfahren (GERBERT 1982)? Auch diese für sich genommen sind u.E. unzureichend, da in Fachtexten ein komplexes Wechselverhältnis zwischen Kommunikationsgegenstand (Themen!), Kommunikationsverfahren und grammatisch-lexikalischen Mitteln unterstellt werden kann. Mit GERBERT plädieren wir für eine "sinnvolle Verflechtung der drei Komponenten", die die Möglichkeit

37 APELT (1974) sieht eine "praktikable Möglichkeit" für Lesekurse darin, daß die Studenten selber Texte vorschlagen: "Nur so kann die Indoktrination, die besonders bei einem Lesekurs für die Sozialwissenschaften schnell gegeben ist, weitgehend vermieden werden" (S. 125). U.E. scheint dieses - zumindest für die Anfangsphase - gar nicht so praktikabel zu sein, da bei einem solchen Verfahren die o.g. Ebenen nicht systematisch erfaßt werden und daher gerade nicht adressaten- und bedarfsorientierte Arbeit geleistet werden kann. Erst zum Schluß eines Lesekurses oder in einem späteren Aufbaukurs bietet das Verfahren wirklich den Vorteil, den es sich eigentlich verspricht, nämlich motivationsfördernd zu sein. Was den Vorwurf der Indoktrination betrifft, so können wir uns dem nicht ganz anschließen: Die Texte werden von den gleichen Fachwissenschaftlern ausgewählt, denen die Studenten ohnehin im Studium "ausgesetzt" sind - Indoktrination könnte vielleicht fortgesetzt, aber nicht durch den FFSU erst geschaffen werden. Außerdem scheint uns der FFSU einer solchen möglichen Indoktrination geradezu entgegenzuwirken, da das erklärte Ziel die Entwicklung einer kritischen Lesehaltung ist. Andere Möglichkeiten der Textauswahl sollten jedoch gesucht und erprobt werden: Diskussionswürdig erscheint uns der Vorschlag bei KUMMER (1983), bei der Textauswahl von den sozialen Problemen der Studenten selbst auszugehen und diese durch Texte zu thematisieren. Auch der Vorschlag eines Fachwissenschaftlers verdient Beachtung, der von sozialen Problemfeldern auszugehen vorschlägt, denen wiederum Texte zugeordnet werden sollen.

offen läßt, "die Lehrbuchtexte in thematischer Anordnung zu bieten und die Sprachausbildung weitgehend unter Dominanz der Verfahrensbezogenheit durchzuführen, weil die Übungen zu den Texten das ermöglichen" (S. 106).

Hinsichtlich unserer Zielgruppe gehen wir von drei Voraussetzungen aus, die sich auf den Aufbau des Lehrwerks auswirken:

1. Die Studenten verfügen meist nur über "verschüttete" Deutschkenntnisse, so daß grammatische Strukturen und Lexik zu reaktivieren sind. Außerdem hat eine Überleitung vom Allgemeindeutsch der Schule zum Fachdeutsch der Hochschule zu erfolgen.
2. Die Studenten stehen meist am Anfang des Studiums und verfügen noch nicht über "Techniken des wissenschaftlichen Arbeitens", so daß "study skills" und "reference skills" in das Lehrwerk einbezogen werden müssen.
3. Die Studenten kommen aus verschiedenen sozialwissenschaftlichen Disziplinen in einen Sprachkurs. Aus außerunterrichtlichen Gründen müssen die Zielgruppen von Lehrwerken aber genügend groß sein, so daß die Adressatenorientierung optimiert, aber nicht maximiert werden kann.

Diese Voraussetzungen berücksichtigend kommen wir zu einer Dreiteilung unseres Lehrwerks:

Teil III (etwa 6 Stunden)	SOZIALWISSENSCHAFTLICHE DISZIPLINEN - Denk- und Mitteilungsstrukturen der einzelnen Disziplinen - fachspezifische Lexik
	Staatswissenschaften / Sozialpsychologie / Sozialpolitik / Journalistik / Verwaltungswissenschaften
	fakultativ
Teil II (etwa 30 Stunden)	PROPÄDEUTIK DER SOZIALWISSENSCHAFTEN - lexikalische und syntaktische Mittel - Denk- und Mitteilungsstrukturen der Sozialwissenschaften - wissenschaftlicher Diskurs - Lesestile
Teil I (etwa 20 Stunden)	EINFÜHRUNG IN DIE WISSENSCHAFTLICHE FACHSPRACHE - lexikalische und syntaktische Mittel - "study skills" und reference skills"

AUFBAU DES LEHRWERKS
(Zu jedem Teil stehen den Lernenden ein Text- und Arbeitsbuch zur Verfügung, die im Anhang eine kontrastive Identifikationsgrammatik und ein zweisprachiges Wörterverzeichnis mit dem Fachwortschatz enthalten.)

Teil I besteht aus 6 Kapiteln und führt in die wissenschaftliche Fachsprache ein [38]. Vom Ansatz her ist Teil I eher verfahrens- als themenorientiert: Im 1. Kapitel werden Titelangaben von Büchern und Zeitschriften, im 2. Kapitel Inhaltsverzeichnisse derselben gelesen. Im 3. Kapitel werden Klappentexte und Abstracts eingesetzt und im 4. Kapitel außersprachliche Kommunikationsverfahren (Tabellen, Diagramme usw.). Erst im 5. (Grundbegriffe der Sozialwissenschaften) und im 6. Kapitel (Grundbegriffe des Rechts) werden Themen aufgegriffen, d.h. zunächst nur Begriffe definiert oder beschrieben.

Teil II ist eine Synthese von thematischer und verfahrensbezogener Progression. Er besteht aus acht Kapiteln [39], die in Anlehnung an die Propädeutik der Sozialwissenschaften entstanden sind, eine solche aber natürlich nicht ersetzen können. Als Textsorten werden Wörterbuchartikel, Zeitschriftenaufsätze, wissenschaftliche Rezensionen, Lehrbücher und wissenschaftliche Monographien herangezogen. Bei den Kommunikationsverfahren herrschen Beschreiben, Berichten, Definieren, Darstellen usw. vor. Erst gegen Ende des zweiten Teils werden komplexere Kommunikationsverfahren eingeführt.

Teil III schließlich besteht aus fünf Textsammlungen zu verschiedenen Fachbereichen, aus denen die meisten unserer Studenten kommen. Diese Textsammlungen sind fakultativ, und der Student wählt nach eigenem Interesse. An Textsorten finden fast nur noch Zeitschriftenaufsätze und wissenschaftliche Monographien Berücksichtigung. Die Kommunikationsverfahren reichen hier bis hin zum Argumentieren, Beweisen usw. Außer den mehr oder weniger hochspezialisierten Texten enthält Teil III Hinweise auf deutschsprachige Hilfsmittel (Wörterbücher, Handbücher usw.) und wichtige wissenschaftliche Zeitschriften der jeweiligen Fachdisziplinen. Eine selbständige Weiterarbeit der Studenten soll so schon durch das Lehrwerk vorbereitet werden.

Zur Erstellung eines solchen Curriculums sind natürlich genaue (o.g.) Textanalysen erforderlich. Es empfiehlt sich daher, wie bereits oben angemerkt, der Textanalyse von Anfang an genügend Aufmerksamkeit zu widmen und

---

[38] Thematisch ist Teil I so neutral gehalten, daß er als Einführung auch für andere Geisteswissenschaften verwendbar ist.

[39] 1. Zur Arbeitsteilung in den Sozialwissenschaften, 2. Sozialwissenschaftliche Disziplinen, 3. Grundlagen der Politischen Ökonomie, 4. Herrschaft und Ideologie,5. Staat und Verwaltung, 6. Soziale Bewegung und Theorie der Gesellschaft, 7. Sozialphilosophie, 8. Methoden der empirischen Sozialforschung. Jedem Kapitel sind mehrere - z.T. konträre - Texte zugeordnet; im Anhang befindet sich ein Überblick über Leben und Werk der Autoren, von denen Texte verwendet werden.

für jeden Text eine "Textbegleitkarte" (KOWALKE) zu erstellen. Die Einordnung in das Curriculum und die Bestimmung der Progression wird dadurch enorm erleichtert, wenn nicht überhaupt erst ermöglicht. Außerdem ist Sorgfalt bei der Textanalyse später zur Entwicklung eines Übungssystems wieder von Nutzen. Daß auch die Bestimmung der Progression und des Aufbaus des Lehrwerks letztlich nur unter Einbeziehung von Fachleuten erfolgen kann, ergibt sich bereits aus dem zuvor Ausgeführten.

Übungssysteme

Wir wenden uns nun der eigentlichen Lehrwerkproduktion zu: der Erstellung eines Übungssystems. Auch hier haben wir von der Zielgruppe und den Lernzielen auszugehen und die Textebene wiederum besonders zu berücksichtigen: Das Übungssystem soll adressatenadäquat sein und die "Arbeitstechniken des Wissenschaftsbetriebes" (WILLE 1978) zugrunde legen.

Vorschläge für fachsprachliche Übungstypologien wurden mehrfach unterbreitet (BECKER 1973 und 1981, BUHLMANN 1982a, BERNSTEIN 1981 und jüngst BEIER/MÖHN 1983). Eine Auseinandersetzung mit diesen würde den Rahmen dieses Berichts sprengen, und wir beschränken uns im folgenden darauf, ein Übungssystem für unser spezielles Bedingungsgefüge zu begründen [40].

Übungen benötigen wir für unser Lehrwerk auf fünf Ebenen:

1. Übungen zur Entwicklung der Techniken des wissenschaftlichen Arbeitens (möglichst kontrastiv!),
2. Übungen zur Entschlüsselung der Textebene (Erkennen der inhaltlichen und logischen Gliederung, der Kommunikationsverfahren und Textsorten),
3. Übungen zur Auseinandersetzung mit dem Inhalt von Texten (kritische Lesehaltung!).
4. Übungen zur Entschlüsselung auf Satzebene (Grammatik!),
5. Übungen zur Entschlüsselung auf Wortebene (Lexik!).

Übungen zu diesen Ebenen stehen zu den o.g. Lernzielen in einem funktionalen Verhältnis. In Teil I bilden die Ebenen 1, 4 und 5 den Schwerpunkt, da erst Voraussetzungen für die Arbeit mit Texten geschaffen werden müssen. Teil II orientiert besonders auf die Ebenen 2 und 3 und hat folgenden Übungsablauf:

1. Phase: Übungen zur Vorentlastung (vor dem Lesen)
   - Bestimmung der Leseinteressen

[40] Wir beziehen uns dabei auf die bereits zitierte Literatur. Vgl. außerdem die Ausführungen bei NEUNER (1981), LÖSCHMANN/PETZSCHLER (1976) und VITLIN (1983).

   - Aktivierung des Vorwissens und der Erwartungen

2. Phase: Übungen zum Grobverständnis (zum ersten Lesen)
   - Erkennen der Schlüsselwörter und der inhaltlichen Gliederung
   - Identifizieren der Mitteilungsstrukturen und Kommunikationsverfahren

3. Phase: Übungen zum Detailverständnis (nach dem Lesen)
   - Entschlüsseln der verständnisrelevanten Lexik und grammatischen Strukturen
   - Erkennen des wissenschaftlichen Diskurses
   - Auseinandersetzung mit dem Inhalt (Vergleichen, Einordnen, Diskutieren, Kritisieren).

Übungen auf den Ebenen 4 und 5 sind in Teil II nur instrumental und orientieren die Studenten immer auf den Kontext. In Teil III geht es hauptsächlich um eine inhaltliche Auseinandersetzung mit Texten, so daß die Ebene 3 präferiert wird.

Wir verzichten hier auf Belege für einzelne Übungstypen zu den verschiedenen Ebenen und verweisen auf die vielen Beispiele, die in der bereits zitierten Literatur genannt werden [41].

Was unser Verfahren bei Übungen zur Bewältigung von Grammatik und Lexik angeht, sei daran erinnert, daß wir grundsätzlich von der Textebene ausgehen und auch in Übungen auf diese zielen, so daß entsprechende "funktionale" Übungen zu entwickeln waren. Ein Übersetzen längerer Textstellen versuchen wir durch geeignete Übungen und Zeitvorgaben zu verhindern, wenngleich wir - zumindest in Teil I - Übersetzungsübungen zur Automatisierung von (aus dem finnischen betrachtet) schwierigen Strukturen nutzen [42]. Die (zum Leseverstehen notwendige Grammatik vermitteln wir kontrastiv und im Sinne einer

[41] Besonders nützlich scheinen uns die Übungsvorschläge bei BUHLMANN (1981 a und b) zu sein, die die Textebene konsequent einbeziehen und hinreichend berücksichtigen.

[42] Vgl. hierzu die grundsätzlichen Ausführungen bei BAZIJEV/TROJANSKAJA (1975), die von drei Aspekten des Übersetzens ausgehen: 1) Übersetzen als Ziel des Unterrichts, 2) Übersetzen als Kontrollform im Leseunterricht und 3) Übersetzen als Mittel zum Textverstehen (s.87). Die Autoren folgern, daß das Übersetzen in der Anfangsphase gerade ein geeigneter Ansatz ist, um zum übersetzungsfreien Leseverstehen zu führen: in der Anfangsphase soll schriftlich übersetzt werden (auf der Grundlage einer vergleichenden Analyse), in der zweiten Phase wird nur noch mündlich übersetzt und in der dritten Phase kann auf das Übersetzen ganz verzichtet werden. In Bezug auf die finnischen Verhältnisse scheint uns dieser Ansatz besonders geeignet, wenn sich die Übersetzungsübungen auf die Anfangsphase und ausgewählte Strukturen beschränken. Kritische Hinweise zum Übersetzen im FFSU finden sich bei NARDON (1983), SCHWIRZ (1983) und ARMALEO-POPPER (1976).

Identifikationsgrammatik (BERNSTEIN 1981, 1983 und 1984): Die Studenten sollen ausgewählte grammatische Strukturen in ihrer Funktion im Text kennen und erkennen (=identifizieren) können. Dabei gehen wir davon aus, daß es für die einzelnen (einzuübenden) Kommunikationsverfahren bestimmte grammatische Signale [43] gibt, durch deren Kenntnis den Studenten das "Sich-Zurechtfinden-im-Text" erleichtert wird. Besonders hier sehen wir ein Feld, das in Zukunft weiter zu bearbeiten wäre, da wir, was eine solche Identifikationsgrammatik betrifft, erst am Anfang stehen [44].

Zum Schluß sei noch darauf hingewiesen, daß auch bei der Erstellung des Übungssystems (insbesondere natürlich bei Übungen auf der Text- und Inhaltsebene) die Zusammenarbeit mit Fachwissenschaftlern unerläßlich ist [45], da Übungen nicht nur adressatenadäquat, sondern auch vom Standpunkt des Faches aus betrachtet authentisch sein müssen.

****

In unseren Anmerkungen zu einer textorientierten Lehrwerkplanung für den FFSU haben wir gezeigt, daß eine konsequente Hinwendung zum Fachtext auf allen Entscheidungsebenen möglich und sinnvoll ist, der Fachtext also Ausgangs- und Endpunkt im FFSU sein kann. Voraussetzung für ein solches Vorgehen sind die interidisziplinäre Zusammenarbeit vor allem mit Fachwissenschaftlern der jeweiligen Disziplinen und umfangreiche Arbeiten zur Textanalyse. Eine Beurteilung nach den Kriterien des "Hamburger Gutachtens" und die Unterrichtsdurchführung haben nun ein so entstandenes Lehrwerk zu evaluieren und erforderlichenfalls zu korrigieren. Aber auch in dieser letzten Phase gilt es verstärkt pragmatische Kategorien heranzuziehen, die zur Evaluierung des Hauptlernziels (= Arbeit mit Texten) besser als linguistische Kategorien geeignet zu sein scheinen.

---

[43] Z.B. Signale für Definieren, Beschreiben usw.; mit ... ist gemeint, unter ... versteht man, ... ist zu verstehen als.

[44] Vgl. auch MATTUSCH (1984) "Ein Fernziel einer kommunikativ-funktionalen Betrachtung von Fachsprachen ... ist die kommunikativ-funktionale, d.h. funktional-semantische und kommunikative, Darstellung der Grammatik auch für den fachsprachlichen Kommunikationsbereich" (S. 73).

[45] Wir müssen hier darauf verzichten, detailliert auf die beabsichtigten Unterrichtsverfahren einzugehen. Dazu nur kurz soviel, daß wir dabei die Arbeitsweise der Fächer berücksichtigen und auf eine Form des "solidarischen Lernens" orientieren, bei der die Kleingruppenarbeit im Mittelpunkt steht. Desweiteren klammern wir hier das Problem einer "fachsprachlichen Landeskunde" aus, die natürlich im Bereich der Sozialwissenschaften eine wichtige Rolle spielen müßte und in unserem Lehrwerk auch Berücksichtigung findet. Verwiesen sei in Bezug auf kulturspezifische Aspekte der Wissenschaftssprachen auf WEIN (1960), HARTMANN (1960) und KUSSMAUL (1978). Vgl. au erdem die didaktischen Hinweise bei BEIER/MÖHN (1981), AHRENS (1981), ZIMMERMAN (1984) und NAUCK (1983).

LITERATUR

Ahrens, Renate E.: Lesefähigkeit in deutscher Fachsprache. Ein Vorschlag zum Kursaufbau. In: Fachsprache, 3-4, 1981, S. 150-168.

Andrä, Helgard: Zur Notwendigkeit und zum Inhalt einer Textanalyse für fremdsprachenmethodische Zwecke. In: Wiss. Zeitschrift der Karl-Marx-Universität Leipzig, Ges. u. Sprachwiss. R., 31, 1982, 1, S. 79-90.

Apelt, Hans-Peter: Zur Frage der Textauswahl in einem Lesekurs für die Sozialwissenschaften. In: Zielsprache Deutsch, 3, 1974.

Armaleo-Popper, Lore: Lesekurse für Anfänger - Fachbereich Psychologie. Methodik und Texte. In: Zielsprache Deutsch, 4, 1976.

Babajlova, A.E./Kicaeva, L.M.: Der Einfluß der inneren und äußeren Absatzstruktur auf das Verstehen fremdsprachiger Lehrtexte. In: Deutsch als Fremdsprache, 2, 1984, S. 106-110.

Barten, Herbert: Ideologiegebundenheit als übergreifendes Merkmal allgemeinsprachlichen und fachsprachlichen Wortschatzes der Gesellschaftswissenschaften. In: cizi jazyky ve skole. XXVII-1983/84, 5.

Becker, Norbert: Versuch einer Übungstypologie für den fachbezogenen Fremdsprachenunterricht. In: Zielsprache Deutsch, 1. 1973, S. 1-10.

Becker, Norbert: In der fachsprachlichen Didaktik ist der fachneutrale Vorkurs ein Umweg. In: Zielsprache Deutsch, 4, 1974, S. 175-178.

Becker, Norbert: Bedarfs- und Kriterienraster für fachsprachliche Lehrwerke. In: Materialien Deutsch als Fremdsprache, Heft 18, Regensburg 1981.

Beier, Rudolf/Möhn, Dieter: Fachsprachlicher Deutschunterricht. Vorüberlegungen zu einem "Hamburger Gutachten" über fachsprachliche Lehr- und Lernmaterialien des Deutschen als Fremdsprache. In: Fachsprache, 3-4. 1981.

Beier, Rudolf/Möhn, Dieter: Merkmale fachsprachlicher Übungen. Beschreibungskriterien für das "Hamburger Gutachten". In: Jahrbuch Deutsch als Fremdsprache, Band 9, München 1983.

Beier, Rudolf/Möhn, Dieter: Fachtexte in fachsprachlichen Lehr- und Lernmaterialien für den fremdsprachlichen Unterricht - Überlegungen zu ihrer Beschreibung und Bewertung. In: Fachsprache, 3-4, 1984, S. 89-115.

Beisbart, Ortwin/Dobnig-Jülch, Edeltraud/Eroms, Hans-Werner/Koss, Gerhard: Textlinguistik und ihre Didaktik, Donauwörth 1976.

Bernstein, Wolf Z.: Leseunterricht und Identifikationsgrammatik. In: Lebende Sprachen, 1, 1981, S. 15-18.

Bernstein, Wolf Z.: Die Ambiguität als verständnishemmender Faktor im Leseunterricht. In: Lebende Sprachen, 4. 1983, S. 149-154.

Bernstein, Wolf Z.: Klassifizierung der Lernschwierigkeiten im Leseverständnis und einige Hinweise zu deren Überwindung. In: Zielsprache Deutsch, 2, 1984

Boeck, Wolfgang/Kirsten, Hans: Zur kommunikativ-funktionalen Sprachbetrachtung. In: Wiss. Zeitschrift der Martin-Luther-Universität Halle-Wittenberg, 5, 1975, S. 55-68.

Buhlmann, Rosemarie: Hinführung zur mathematisch-naturwissenschaftlichen Fachsprache. Ein Beitrag zur Lösung von Problemen bei der Vermittlung mathematisch-naturwissenschaftlicher Fachsprache an sprachlich-heterogenen Adressaten. In: Materialien Deutsch als Fremdsprache, Heft 11, Regensburg 1978.

Buhlmann, Rosemarie: Das Lesen von Fachtexten. In: Das Lesen in der Fremdsprache. Beiträge eines Werkstattgesprächs des Goethe-Instituts New York und des ACTFL. New York vom 25. bis 28.9.1979. Herausgegeben von Helm von Faber und Manfred Heid, München 1981a.

Buhlmann, Rosemarie: Texte Verstehen. In: Schriftenreihe des türkisch-deutschen Kulturinstituts Istanbul, 2, 1981b, S. 51-96.

Buhlmann, Rosemarie: Analyse und Beurteilung fachsprachlicher Lehrwerke: Kriterien und ihre Problematik. In: Krumm, Hans-Jürgen (Hrsg.): Lehrwerkforschung - Lehrwerkkritik Deutsch als Fremdsprache, München 1982a.

Buhlmann, Rosemarie: Sprachliche Handlungsfähigkeit im Fach als Ziel der Fachsprachendidaktik. In: Deutsch als Fachsprache, Poznan 1982b.

Buhlmann, Rosemarie: Die Problematik des Fachsprachenunterrichts im Bereich Deutsch als Fremdsprache. In: Kielikeskusuutisia-Language Centre News, 4. 1983.

Ebermann, Joachim/Gebhardt, Kurt/Schindhelm, Waldemar/Zuber, Gerhard: Probleme und Erfahrungen bei der Gestaltung von Lehrbüchern für die fachsprachliche Ausbildung. In: Wiss. Zeitschrift der TU Dresden, 23, 1974, 3/4, S. 677-682.

Ehnert, Rolf: Fachtexte verstehen. Probleme - Kleinere Arbeitsvorhaben - vorgestellt. In: Bielefelder Beiträge zur Sprachlehrforschung, 7, 1976, S. 31-39.

Fachsprachen. Herausgegeben von Walther von Hahn, Darmstadt 1981.

Fluck, Hans-Rüdiger: Fachsprachen. Einführung und Bibliographie, München 1976, 2. durchgesehene und erweiterte Auflage 1980.

Fluck, Hans-Rüdiger: Deutsch als Fachsprache in der Postgraduiertenausbildung an der Tongji-Universität, Shanghai. In: Muttersprache, 94, 1983/84, S. 161-174.

Gerbert, Manfred: Elemente einer Lehrbuchtheorie für die Fremdsprachenausbildung. In: Wissenschaftliche Zeitschrift der TU Dresden, 31, 1982, 5, S. 105-108.

Gläser, Rosemarie: Die funktionalstilistische Komponente in der fachsprachlichen Forschung und Lehre. In: Wiss. Zeitschrift der Humboldt-Universität, Ges. u. Sprachwiss. Reihe, XXVII, 1978, 4, S. 463-465.

Grawert, Ursula: Zum Beweisen in argumentierenden Texten aus dem Bereich der Publizistik aufgearbeitet für den Sprachunterricht Deutsch als Fremdsprache. Dissertation (Humboldt-Universität zu Berlin), Berlin 1982.

Grosso, Ingrid: Lesekurse für Geisteswissenschaftler am Goethe-Institut Turin 1976-1980. In: Materialien Deutsch als Fremdsprache, Heft 18, Regensburg 1981.

Hartmann, Hans: Die Struktur der indogermanischen Sprachen und die Entstehung der Wissenschaft. In: Sprache und Wissenschaft, a.a.O.

Herrmann, Karin: Konzeption und Durchführung von Fachlesekursen für Geisteswissenschaftler am Goethe-Institut Turin, 1976-1978. In: Lesen in der Fremdsprache, a.a.O.

Hoffmann, Lothar: Kommunikationsmittel Fachsprache. Eine Einführung, Berlin 1976.

Hoffmann, Lothar: Fachtextlinguistik. In: Fachsprache, 2, 1983, S. 57-67.

Hoffmann, Lothar: Fachwortschatz - Grundwortschatz - Minimum. In: Deutsch als Fremdsprache, 4, 1984a.

Hoffmann, Lothar: Vom Fachtext zur Fachtextsorte. In: Deutsch als Fremdsprache, 6, 1984b.

Janda, Josef W./Sprissler, Manfred: Bedarfsanalyse und Textauswahl. Zur Entstehung eines fachspezifischen Analysecorpus. In: Bielefelder Beiträge zur Sprachlehrforschung, 9, 1977.

John, Tim/Davies, Florence: Text as a Vehicle for Information: the Classroom Use of Written Texts in Teaching Reading in a Foreign Language. In: Reading in a Foreign Language, Vol. 1, No. 1, March 1983.

Kamprad, Walter: Über die Struktur von Absätzen gesellschaftswissenschaftlicher Texte und zu Modellen für die Textarbeit. In: Deutsch als Fremdsprache, 4, 1975, S. 223-234.

Kamprad, Walter: Zur Auswahl und Gestaltung von Fachtexten im Deutschunterricht für Fortgeschrittene. In: Deutsch als Fremdsprache, 1, 1971, S. 47-51.

Karajoli, Edeltraud: Strategien des fremdsprachigen Leselernprozesses. Fehleranalytische Auswertung eines psycholinguistischen Experiments, Berlin, Techn. Univ. FB 1, Dissertation 1981.

Keiser, Robert: Landes-, Fremd- und Fachsprachen in der schweizerischen Wirtschaft. In: Kelz, Heinrich P., a.a.O.

Kelz, Heinrich P.: Fachsprache 1: Sprachanalyse und Vermittlungsmethoden. Dokumentation einer Tagung der Otto-Benecke-Stiftung, Bonn 1983.

Köhler, Claus: Bemerkungen zum Problem eines variabel einsetzbaren Lehrbuchs für den fachbezogenen Fortgeschrittenenunterricht. In: Deutsch als Fremdsprache, 5, 1980, S. 304-309.

Köhler, Claus: Zur Didaktik des Deutschen als Fachsprache - eine spezielle Modifikation fremdsprachendidaktischer Fragestellung. In: Schröder, Hartmut (Hrsg.): Beiträge zu einer Linguistik und Didaktik des Deutschen als Fremdsprache, Jyväskylä 1985

Kowalke, Hermann: Zur Methodologie der Erarbeitung fachsprachlicher Lehrbücher. In: Wiss. Zeitschrift der Univ. Rostock, Ges. und Sprachwiss. Reihe, 2, 1973.

Kummer, Manfred: "Soziologendeutsch" - Planung und Vorarbeiten für fachspezifische Sprachkurse. In: Kelz, Heinrich P., a.a.O.

Kussmaul, Paul: Kommunikationskonventionen in Textsorten am Beispiel deutscher und englischer geisteswissenschaftlicher Abhandlungen. Ein Beitrag zur deutsch-englischen Übersetzungstechnik. In: Lebende Sprachen, 2, 1978, S. 54-58.

Langer, Inghard/Schulz V. Thun, Friedemann/Tausch, Reinhard: Verständlichkeit in Schule, Verwaltung, Politik und Wissenschaft, München 1974.

Löschmann, Martin/Petzschler, Hermann: Übungsgestaltung zum verstehenden Hören und Lesen, Leipzig 1976.

Mattusch, Hans-Jürgen: Zur Betrachtung der Aufforderungsmodalität unter kommunikativ-funktionalem Aspekt - untersucht an einigen naturwissenschaftlichen Fachsprachen des Russischen. In: Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung, 37, 1984, 1.

Möhn, Dieter: Fachtexte. In: Fachsprache, 2, 1983, S. 50-51.

Nardon, Erika: Fachsprache als Fremdsprache an italienischen Hochschulen. In: Punkt Absatz. Zeitschrift für DaF an der italienischen Hochschule, 2, 1983.

Nauck, Hans-Joachim: Schwerpunktverlagerungen im Fremdsprachenunterricht. In: Wirtschaft und Gesellschaft im Unterricht, 8, 1983, S. 116-119.

Neuf-Münkel, Gabriele: Didaktisierung von Texten aus dem Bereich der Gesellschaftswissenschaften. In: Materialien Deutsch als Fremdsprache, Heft 19, Regensburg 1982.

Neuner, Gerhard/Krüger, Michael/Grewer, Ulrich: Ubungstypologie zum kommunikativen Deutschunterricht, Berlin und München 1981.

Nübold, Peter: Fachspezifische Englischkenntnisse für das Studium der Ingenieurwissenschaften: Bedarfsermittlung und Kurskonzeption. In: Gnutzmann/Turner: Fachsprachen und ihre Anwendung, Tübingen 1980.

Ortmann, Hanni/Sternagel, Helga: Zur Charakterisierung von Fachtexten für Könnensentwicklung im fachbezogenen Fremdsprachenunterricht. In: Wiss. Zeitschrift der Humboldt-Universität zu Berlin, Ges. u. Sprachwiss. Reihe, XXVII, 1978, 4, S. 493-496.

Petöfi, Jãnos S.: Einige allgemeine Aspekte der Analyse und Beschreibung wissenschaftssprachlicher Texte. In: Bungarten, Theo (Hrsg.), Wissenschaftssprache. Beiträge zur Methodologie, theoretischen Forschung und Deskription, München 1981.

Poetzelberger, Hans A.: Fachbezogene und tätigkeitsspezifische Sprachkompetenz. In: Kelz, Heinrich P., a.a.O.

Rothacker, Erich: Die Sprache der Geisteswissenschaften. In: Sprache und Wissenschaft, a.a.O.

Schilling, Irmgard: Sachverhalte und Syntax beim Erwerb fachorientierter Fremdsprachenkenntnisse. In: Deutsch als Fremdsprache, 3, 1973, S. 176-182.

Schleusener, Klaus: Probleme der Integration von Fachsprache und Umgangssprache. Werkstattbericht. In: Fachsprache, 4, 1982.

Schlieben-Lange, Brigitte/Kreuzer, Helmut: Probleme und Perspektiven der Fachsprachen- und Fachliteraturforschung. Zur Einleitung. In: Zeitschrift für Literaturwissenschaft und Linguistik, 51/52, 1983, S. 7-26.

Schmidt, Wilhelm: Was sind Kommunikationsverfahren? In: Fremdsprachenunterricht, 2/3, 1980, S. 128-135.

Schröder, Hartmut: Deutsch als Fremdsprache für Nichtphilologen an finnischen Hochschulen. In: Informationen Deutsch als Fremdsprache, 6, 1983/84, S. 39-46.

Schwirtz, Käthe: Methodisch relevante Beziehungen der Sprachtätigkeit Lesen beim Ubersetzen. In: Beiträge zur Methodik des fachspezifischen Fremdsprachenunterrichts (III), Greifswald 1983.

Selle, Sigrid: Die französische Sprache im wissenschaftlichen Text - einige Aspekte fachsprachlicher Forschung. In: Fachsprache, 2, 1983, S. 68-73.

Seltmann. Wolfgang: Die Leseleistung im fremdsprachigen Fachtext und ihre Entfaltung. Dissertation B (Karl-Marx-Universität Leipzig), Leipzig 1978.

Spillner, Bernd: Methodische Aufgaben der Fachsprachenforschung und ihre Konsequenzen für den Fremdsprachenunterricht. In: Kelz, Heinrich P., a.a.O.

Spitzbardt, Harry: Zur Integration der fachsprachlichen Komponente im Deutschunterricht für Ausländer. In: Deutsch als Fremdsprache, 6, 1983.

Sprache und Wissenschaft. Vorträge gehalten auf der Tagung der Joachim Jungius-Gesellschaft der Wissenschaften, Hamburg, am 29. und 30. Oktober 1959, Göttingen 1960.

Sprissler, Manfred: Fachspezifische Sprachlehre und Fachsprachenlinguistik. In: Bielefelder Beiträge zur Sprachlehrforschung, 9, 1977.

Stanley, Rose Mary: The Recognition of Macrostructur: A Pilot Study. In: Reading in a Foreign Language, Vol. 2, No. 2, Spring 1984, S. 156-168.

Uesseler, Manfred: Probleme bei der Abgrenzung zwischen Fachsprache und Allgemeinsprache im Bereich der Gesellschaftswissenschaften. In: Deutsch als Fachsprache, Poznan 1982.

Vitlin, Z.L.: Leseziele im Fremdsprachenunterricht für Erwachsene und Uberwindung von Verstehensschwierigkeiten beim Lesen deutscher Originaltexte. In: Deutsch als Fremdsprache, 6, 1983.

Wein, Hermann: Sprache und Wissenschaft. In: Sprache und Wissenschaft, a.a.O.

Wille, Konrad: Auswahlkriterien und Ubungstypologie eines Unterrichtsprogramms "Wissenschaftssprache". In: Materialien Deutsch als Fremdsprache, Heft 11, Regensburg 1978.

Williams, Ray: Teaching the Recogniton of Cohesive Ties in Reading in a Foreign Language. In: Reading in a Foreign Language, Vol. 1, No. 1, March 1983, S. 35-52.

Zimmermann, Klaus: Einige Hypothesen bezüglich Leseverstehen im L2-Erwerb. In: Informationen Deutsch als Fremdsprache, 1, 1984/85.

Академия наук СССР/Кафедра иностранных языков: Обучение чтению научного текста на иностранном языке. Москва 1975.

Базиев, А.Т./Троянская, Е.С.: Перевод как один из эффективных способов обучения чтению и пониманию научной и технической литературы. In: ebenda.

Мету, Н.А./Митрофанова, О.Д./Одинцова, Т.Б.: Структура научного текста и обучение монологической речи. Москва 1981.

SUBSCRIPTION FORM

FINLANCE
The Director, Language Centre for Finnish Universities
University of Jyväskylä
SF-40100 Jyväskylä 10
Finland

Please enter my subscription to FINLANCE (indicate your choice):

☐ standing order, at 20 Fmk per issue
☐ next issue only, at 20 Fmk

Each issue will be accompanied by an invoice, i.e. no cheques should be sent in advance. A subscription form will be printed at the back of each issue, for the convenience of those who wish to subscribe to only one issue at a time.

Additionally, I wish to order the following issues:

	Price	Number of copies
☐ FINLANCE, volume I	20 Fmk	______
☐ FINLANCE, volume II	20 Fmk	______
☐ FINLANCE, volume III	20 Fmk	______
☐ Teaching and testing communicative competence (1976)	8 Fmk	______
☐ Focus on spoken language (1978)	10 Fmk	______
☐ The language laboratory: methods and materials (1979)	12 Fmk	______

Name ______________________________

Address ______________________________

____

____

Signature ______________________________